



1 of 1

Download Print Save to PDF Add to List Create bibliography

 Page could not be loaded  
Please try again later

References (49)

[View in search results format >](#)

All Export Print E-mail Save to PDF Create bibliography

1 Chithaluru, P., Singh, A., Mahmoud, M.S., Kumar, S., Mazón, J.L.V., Alkhayat, A., Anand, D.  
[An enhanced opportunistic rank-based parent node selection for sustainable & smart IoT networks](#)

Cited by 16 documents

[Dynamic clustering optimization for energy efficient IoT Network: A simple constrastive graph approach](#)

Raj, R.S. , Hema, L.K.  
(2025) *Expert Systems with Applications*

[Industrial mechanical equipment fault detection and high-performance data analysis technology based on the Internet of Things](#)

Ding, D.  
(2024) *Intelligent Decision Technologies*

[Energy efficient data communication for WSN based resource constrained IoT devices](#)

Hudda, S. , Haribabu, K. , Barnwal, R.  
(2024) *Internet of Things (Netherlands)*

[View all 16 citing documents](#)

 View PDF

Download full issue

## Outline

Abstract

Keywords

1. Introduction and literature

2. Basic background subtraction algorithm

3. Background modelling

4. Implementation of modified GMM based obje...


5. Result and discussion

6. Conclusion and future scope

Declaration of competing interest

Data availability

References

Show full outline 

Cited by (10)





## Measurement: Sensors



Volume 30, December 2023, 100898



# Moving object detection using modified GMM based background subtraction

S. Rakesh <sup>a</sup>, Nagaratna P. Hegde <sup>b</sup>, M. Venu Gopalachari <sup>c</sup>, D. Jayaram <sup>c</sup>, Bhukya Madhu <sup>d</sup>, Mohd Abdul Hameed <sup>a</sup>, Ramdas Vankdothu <sup>e</sup>  , L.K. Suresh Kumar <sup>a</sup>

Show more 
 Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.measen.2023.100898> 
Get rights and content Under a Creative Commons [license](#) 
 open access

## Abstract

Academics have become increasingly interested in creating cutting-edge technologies to enhance Intelligent Video Surveillance (IVS) performance in terms of accuracy, speed, complexity, and deployment. It has been noted that precise object detection is the only

Part of special issue 

### Smart Sensor Technology in Renewable Energy Applications

Edited by Jasmine Gnana Malar PSN College of Engineering and Technology, Tirunelveli, Tamil Nadu, India, Kumara Swamidhas L.A. Indian Institute of Technology Dhanbad, Jharkhand, India, Siamak Hoseinzadeh Sapienza Università di Roma, Rome, Italy, Jafar A. Alzubi School of Engineering, Al-Balqa` Applied University, Jordan



View special issue


Recommended articles 

### Motion and appearance based background subtraction for freely moving cameras

Signal Processing: Image Communication, Volume 75, ...  
Hasan Sajid, ..., Nathan Jacobs

 View PDF

Fusion representation learning for

 FEEDBACK



1 of 1

[Download](#) [Print](#) [Save to PDF](#) [Add to List](#) [Create bibliography](#)

 **Page could not be loaded**  
Please reload to try again

[Reload page](#)

References (23)

[View in search results format >](#)

All [Export](#) [Print](#) [E-mail](#) [Save to PDF](#) [Create bibliography](#)

1 Ahmad, R., Wazirali, R., Abu-Ain, T.  
[Machine Learning for Wireless Sensor Networks Security: An Overview of](#)

Cited by 3 documents

[Multi-view consistent generative adversarial network for enhancing intrusion detection with prevention systems in mobile ad hoc networks against security attacks](#)

Rajkumar, M. , Karthika, J. , Abinayaa, S.S. (2025) *Computers and Security*

[Enhancing intrusion detection in MANETs with blockchain-based trust management and enhanced GRU model](#)

Krishna, E.S.P. , Sandeep, D. , Kocherla, R. (2025) *Peer-to-Peer Networking and Applications*


[A secure routing and black hole attack detection system using coot Chimp Optimization Algorithm-based Deep Q Network in MANET](#)

D, S. , PH, L. (2025) *Computers and Security*



1 of 1

Download Print Save to PDF Add to List Create bibliography

 Page could not be loaded  
Please reload to try again

[Reload page](#)

References (23)

[View in search results format >](#)

All | [Export](#) [Print](#) [E-mail](#) [Save to PDF](#) [Create bibliography](#)

1 Ahmad, R., Wazirali, R., Abu-Ain, T.  
[Machine Learning for Wireless Sensor Networks Security: An Overview of](#)

Cited by 3 documents

Multi-view consistent generative adversarial network for enhancing intrusion detection with prevention systems in mobile ad hoc networks against security attacks

Rajkumar, M. , Karthika, J. , Abinayaa, S.S. (2025) *Computers and Security*

Enhancing intrusion detection in MANETs with blockchain-based trust management and enhanced GRU model

Krishna, E.S.P. , Sandeep, D. , Kocherla, R. (2025) *Peer-to-Peer Networking and Applications*


A secure routing and black hole attack detection system using coot Chimp Optimization Algorithm-based Deep Q Network in MANET

D, S. , PH, L. (2025) *Computers and Security*



1 of 1

Download Print Save to PDF Add to List Create bibliography

 Page could not be loaded  
Please reload to try again

[Reload page](#)

References (38)

[View in search results format >](#)

All    [Export](#)    [Print](#)    [E-mail](#)    [Save to PDF](#)    [Create bibliography](#)

1 Stamenovic, M., Schick, S., Luo, J.  
[Machine Identification of High Impact Research through Text and Image Analysis](#)

Cited by 6 documents

[Integrated normal discriminant analysis in mapreduce for diabetic chronic disease prediction using bivariant deep neural networks](#)

Ramani, R. , Dhinakaran, D. , Edwin Raja, S.  
(2024) *International Journal of Information Technology (Singapore)*

[Boosting interclass boundary preservation \(BIBP\): a KD-tree enhanced data reduction algorithm](#)

Fuangkhon, P.  
(2024) *International Journal of Information Technology (Singapore)*

[Cryptanalysis and security evaluation of optimized algorithms for image encryption in deep optimal network](#)

Manoharan, S.N.  
(2024) *International Journal of Information*



1 of 1

Download Print Save to PDF Add to List Create bibliography

 Page could not be loaded  
Please reload to try again

[Reload page](#)

Cited by 0 documents

Inform me when this document is cited in Scopus:

[Set citation alert >](#)

Related documents

[Laddering vision foundation model for remote sensing image change detection](#)  
Liu, Y. , Zhou, G.  
(2024) *Journal of Applied Remote Sensing*

[Risk Analysis of Electricity Balance in New Power System: A Case Study of Yunnan in China](#)  
Xie, M. , Liu, S. , Cai, H.  
(2023) *Proceedings - 2023 International Conference on Power System Technology: Technological Advancements for the*

References (20)

[View in search results format >](#)

All Export Print E-mail Save to PDF Create bibliography

1 Sirmacek, B., Vinuesa, R.  
[Remote sensing and AI for building climate adaptation applications](#)



1 of 1

Download Print Save to PDF Add to List Create bibliography

 Page could not be loaded  
Please reload to try again

[Reload page](#)

References (20)

[View in search results format >](#)

All Export Print E-mail Save to PDF Create bibliography

- 1 Sirmacek, B., Vinuesa, R.  
[Remote sensing and AI for building climate adaptation applications](#)

Cited by 0 documents

Inform me when this document is cited in Scopus:

[Set citation alert >](#)

Related documents

[Laddering vision foundation model for remote sensing image change detection](#)

Liu, Y. , Zhou, G.  
(2024) *Journal of Applied Remote Sensing*

[Risk Analysis of Electricity Balance in New Power System: A Case Study of Yunnan in China](#)

Xie, M. , Liu, S. , Cai, H.  
(2023) *Proceedings - 2023 International Conference on Power System Technology: Technological Advancements for the*



1 of 1

Download Print Save to PDF Add to List Create bibliography

 Page could not be loaded  
Please reload to try again  
[Reload page](#)

Cited by 0 documents

Inform me when this document is cited in Scopus:

[Set citation alert >](#)

Related documents

[Artificial Intelligence Applications and Prospects for the Smart Grid](#)  
Zhao, X. , Guo, Y. , Guo, X. (2023) *Proceedings - 2023 Panda Forum on Power and Energy, PandaFPE 2023*

[Analysis of Short-Term Voltage Stability Influencing Factors and Mechanisms in Low Inertia Active Distribution Networks](#)  
Guo, Q. , Shan, B. , Zheng, W. (2023) *Proceedings - 2023 3rd Power System and Green Energy Conference, PSGEC 2023*

References (25)

[View in search results format >](#)

All | [Export](#) | [Print](#) | [E-mail](#) | [Save to PDF](#) | [Create bibliography](#)


1 Wang, Junfei, Srikantha, Pirathayini  
Fast optimal power flow with guarantees via an unsupervised generative model (2022) *IEEE Transactions on Power Systems*. Cited 5 times.





1 of 1

Download Print Save to PDF Add to List Create bibliography

 Page could not be loaded  
Please reload to try again

[Reload page](#)

References (21)

[View in search results format >](#)

All Export Print E-mail Save to PDF Create bibliography

1 Almaiah, M.A., Ali, A., Hajje, F., Pasha, M.F., Alohali, M.A.  
[A Lightweight Hybrid Deep Learning Privacy Preserving Model for FC-Based Industrial Internet of Medical Things](#)

Cited by 1 document

Modern Diagnostic Imaging Classifications and Risk Factors for 6G-enabled Smart Health Systems

Ramu, K. , Krishnamoorthy, R. , Salim, A. (2023) *Radioelectronics and Communications Systems*

[View details of this citation](#)

Inform me when this document is cited in Scopus:

[Set citation alert >](#)

Related documents

[Smart Healthcare IoT: Deep Learning-Driven Patient Monitoring and Diagnosis](#)

Viswadutt, N.J. , Vemula, D.K. , Shardunya,



1 of 1

Download Print Save to PDF Add to List Create bibliography

 Page could not be loaded  
Please reload to try again

[Reload page](#)

Cited by 0 documents

Inform me when this document is cited in Scopus:

[Set citation alert >](#)

Related research data ?

UnifiedSKG: Unifying and Multi-Tasking Structured Knowledge Grounding with Text-to-Text Language Models

Unknown Repository

ShadowGNN: Graph Projection Neural Network for Text-to-SQL Parser

, et al  
Unknown Repository

References (32)

[View in search results format >](#)

All Export Print E-mail Save to PDF Create bibliography

1 Parikh, P., Chatterjee, O., Jain, M., Harsh, A., Shahani, G., Biswas, R., Arya, K. (1999) *Auto-Query-A simple natural language to SQL query generator for an e-learning platform* New York, Academic Press

All

ADVANCED SEARCH

Journals & Magazines > IEEE Access > Volume: 11 ?

# Ensemble Learning With Tournament Selected Glowworm Swarm Optimization Algorithm for Cyberbullying Detection on Social Media

Publisher: IEEE

Cite This



Ravuri Daniel ; T. Satyanarayana Murthy ; Ch. D. V. P. Kumari ; E. Laxmi Lydia <sup>id</sup> ; Mohamad Khairi Ishak <sup>id</sup> ; Myriam Hadjouni <sup>id</sup> **All Authors**

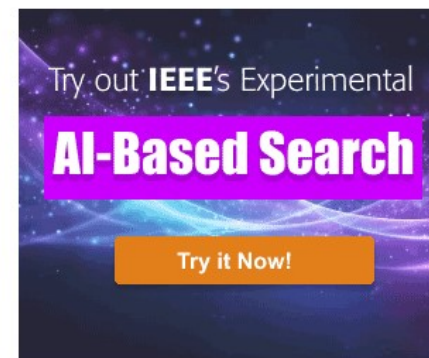
5  
Cites in  
Papers

1033  
Full  
Text Views



Open Access Comment(s)

Under a Creative Commons License



### More Like This

Prediction of heart disease ensemble learning and Particle Swarm Optimization

2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)

PDF

Help

### Abstract

### Abstract:

ng [MathJax]/jax/output/HTML-CSS/fonts/TeX/AMS/Regular/GeneralPunctuation.js N) plays a crucial role to facilitate social connections; but, this social networking media

Document Sections

 View PDF

Download full issue

## Outline

Abstract

Keywords

1. Introduction

2. Literature review

3. The proposed model

4. Results and discussion

5. Conclusion

Data availability statement

Ethics approval

Consent to participate

Informed consent

Declaration of Competing Interest

Data availability

References

Vitae



## Computers and Electrical Engineering

Volume 109, Part A, July 2023, 108765



# Fuzzy wavelet neural network driven vehicle detection on remote sensing imagery ☆

Mohammed Altaf Ahmed <sup>a</sup>✉, Sara A Althubiti <sup>b</sup>✉, Victor Hugo C. de Albuquerque <sup>h</sup>✉, Marcello Carvalho dos Reis <sup>g</sup>✉, Chitra Shashidhar <sup>e,f</sup>, T Satyanarayana Murthy <sup>c</sup>✉, E. Laxmi Lydia <sup>d</sup>✉

Show more ▾

[+ Add to Mendeley](#) [Share](#) [Cite](#)
<https://doi.org/10.1016/j.compeleceng.2023.108765>
[Get rights and content](#)

## Abstract

Remote sensing-based target detection process is applied to spot the targeted objects in remote sensing images (RSIs). However, it is challenging to detect small-sized vehicles in RSIs. The current study designs a Chicken Swarm Optimization with Transfer Learning-Driven Vehicle Detection and Classification on Remote Sensing Imagery (CSOTL-VDPC)

Part of special issue

## Artificial Intelligence-based Sensors for Industrial IoT Applications

Edited by Deepak Gupta, Oscar Castillo

[View special issue](#)

Recommended articles

### A novel framework for crowd counting using video and audio

Computers and Electrical Engineering, Volume 109, Pa...  
Yi Zou, ..., Qing Han
[View PDF](#)

### Novel virtual nasal endoscopy system based on computed tomography scans

Virtual Reality & Intelligent Hardware, Volume 4, Issu...  
Fábio de O. Sousa, ..., Victor Hugo C. de Albuquerque
[View PDF](#)

### An efficient mechanism using IoT and

[FEEDBACK](#)



Computer Systems  
Science and Engineering

[Submit a Paper](#)

[Propose a Special Issue](#)

## Table of Content

- > [Abstract](#)
- > [Introduction](#)
- > [Materials and Methods](#)
- > [Results and Discussion](#)
- > [Conclusion](#)
- > [References](#)

Open Access

ARTICLE

## Artificial Humming Bird Optimization with Siamese Convolutional Neural Network Based Fruit Classification Model

by T. Satyanarayana Murthy<sup>1</sup>, Kollati Vijaya Kumar<sup>2</sup>, Fayadh Alenezi<sup>3</sup>, E. Laxmi Lydia<sup>4</sup>, Gi-Cheon Park<sup>5</sup>, Hyoung-Kyu Song<sup>6</sup>, Gyanendra Prasad Joshi<sup>7</sup>, Hyeonjoon Moon<sup>7,\*</sup>

1 Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana, India

2 Department of Computer Science, College of Computer and Information Sciences, Majmaah University, Al-Majmaah, 11952, Saudi Arabia

3 Department of Electrical Engineering, College of Engineering, Jouf University, Sakaka, Saudi Arabia

4 Department of Computer Science and Engineering, Vignan's Institute of Information Technology, Visakhapatnam, 530049, India

5 Department of International Affairs and Education, Gangseo University, Seoul, 07661, Korea

6 Department of Information and Communication Engineering and Convergence Engineering for Intelligent Drone, Sejong University, Seoul, 05006, Korea

7 Department of Computer Science and Engineering, Sejong University, Seoul, 05006, Korea

\* Corresponding Author: Hyeonjoon Moon. Email: hmoon@sejong.ac.kr

*Computer Systems Science and Engineering* **2023**, 47(2), 1633-1650. <https://doi.org/10.32604/csse.2023.034769>

**Received** 27 July 2022; **Accepted** 26 October 2022; **Issue published** 28 July 2023

[View Full Text](#)

[Download PDF](#)



### Abstract



Downloads



Citation Tools

1070

View

526

Download



### Related articles

Automated and Precise Event Detection Method for Big Data in Biomedical Imaging with Support Vector Machine

Lufeng Yuan, Erlin Yao, Guangming...

Sentiment Analysis System in Big Data Environment



Dr RajaniKanth  
Aluvalu

## Diagnostic structure of visual robotic inundated systems with fuzzy clustering membership correlation

[HTML] from peerj.com

Authors Hariprasath Manoharan, Shitharth Selvarajan, Rajanikanth Aluvalu, Maha Abdelhaq, Raed Alsaqour, Mueen Uddin

Publication date 2023/12/19

Journal PeerJ Computer Science

Volume 9

Pages e1709

Publisher PeerJ Inc.

**Description** The process of using robotic technology to examine underwater systems is still a difficult undertaking because the majority of automated activities lack network connectivity. Therefore, the suggested approach finds the main hole in undersea systems and fills it using robotic automation. In the predicted model, an analytical framework is created to operate the robot within predetermined areas while maximizing communication ranges. Additionally, a clustering algorithm with a fuzzy membership function is implemented, allowing the robots to advance in accordance with predefined clusters and arrive at their starting place within a predetermined amount of time. A cluster node is connected in each clustered region and provides the central control center with the necessary data. The weights are evenly distributed, and the designed robotic system is installed to prevent an uncontrolled operational state. Five different scenarios are used to test and validate the created model, and in each case, the proposed method is found to be superior to the current methodology in terms of range, energy, density, time periods, and total metrics of operation.

Total citations **Cited by 2**





Dr RajaniKanth  
Aluvalu

## Machine learning job failure analysis and prediction model for the cloud environment [HTML] from sciencedirect.com

Authors Harikrishna Bommala, Rajanikanth Aluvalu, Swapna Mudrakola

Publication date 2023/12/1

Journal High-Confidence Computing

Volume 3

Issue 4

Pages 100165

Publisher Elsevier

**Description** Reliable and accessible cloud applications are essential for the future of ubiquitous computing, smart appliances, and electronic health. Owing to the vastness and diversity of the cloud, a most cloud services, both physical and logical services have failed. Using currently accessible traces, we assessed and characterized the behaviors of successful and unsuccessful activities. We devised and implemented a method to forecast which jobs will fail. The proposed method optimizes cloud applications more efficiently in terms of resource usage. Using Google Cluster, Mustang, and Trinity traces, which are publicly available, an in-depth evaluation of the proposed model was conducted. The traces were also fed into several different machine learning models to select the most reliable model. Our efficiency analysis proves that the model performs well in terms of accuracy, F1-score, and recall. Several factors, such as ...

Total citations [Cited by 14](#)



Scholar articles [Machine learning job failure analysis and prediction model for the cloud environment](#)  
H Bommala, R Aluvalu, S Mudrakola - High-Confidence Computing, 2023



Dr RajaniKanth Aluvalu

# A novel artificial intelligence-based predictive analytics technique to detect skin cancer

[HTML] from peerj.com

Authors Prasanalakshmi Balaji, Bui Thanh Hung, Prasun Chakrabarti, Tulika Chakrabarti, Ahmed A Elngar, Rajanikanth Aluvalu

Publication date 2023/5/24

Journal PeerJ Computer Science

Volume 9

Pages e1387

Publisher PeerJ Inc.

Description One of the leading causes of death among people around the world is skin cancer. It is critical to identify and classify skin cancer early to assist patients in taking the right course of action. Additionally, melanoma, one of the main skin cancer illnesses, is curable when detected and treated at an early stage. More than 75% of fatalities worldwide are related to skin cancer. A novel Artificial Golden Eagle-based Random Forest (AGEbRF) is created in this study to predict skin cancer cells at an early stage. Dermoscopic images are used in this instance as the dataset for the system's training. Additionally, the dermoscopic image information is processed using the established AGEbRF function to identify and segment the skin cancer-affected area. Additionally, this approach is simulated using a Python program, and the current research's parameters are assessed against those of earlier studies. The results demonstrate that, compared to other models, the new research model produces better accuracy for predicting skin cancer by segmentation.

Total citations Cited by 7



Scholar articles [A novel artificial intelligence-based predictive analytics technique to detect skin cancer](#)





Dr RajaniKanth  
Aluvalu

## Clustering based EO with MRF technique for effective load balancing in cloud computing

Authors Hanuman Reddy, Amit Lathigara, Rajanikanth Aluvalu

Publication date 2023/5/22

Journal International Journal of Pervasive Computing and Communications

Issue ahead-of-print

Publisher Emerald Publishing Limited

Description Purpose

Cloud computing (CC) refers to the usage of virtualization technology to share computing resources through the internet. Task scheduling (TS) is used to assign computational resources to requests that have a high volume of pending processing. CC relies on load balancing to ensure that resources like servers and virtual machines (VMs) running on real servers share the same amount of load. VMs are an important part of virtualization, where physical servers are transformed into VM and act as physical servers during the process. It is possible that a user's request or data transmission in a cloud data centre may be the reason for the VM to be under or overloaded with data.

Total citations Cited by 5



Scholar articles [Clustering based EO with MRF technique for effective load balancing in cloud computing](#)  
H Reddy, A Lathigara, R Aluvalu - International Journal of Pervasive Computing and ..., 2023  
[Cited by 5](#) [Related articles](#) [All 3 versions](#)



Dr RajaniKanth  
Aluvalu

## Energy optimization in path arbitrary wireless sensor network

Authors Rajanikanth Aluvalu B. Harish Goud, T. N. Shankar, Basant Sah

Publication date 2023/3/16

Journal Expert Systems

Publisher wiley

DOI: [10.14569/IJACSA.2024.01506156](https://doi.org/10.14569/IJACSA.2024.01506156)

## Text Extraction and Translation Through Lip Reading using Deep Learning

[PDF](#)

 Author 1: Sai Teja Krithik Putcha  Author 2: Yelagandula Sai Venkata Rajam  Author 3: K. Sugamya  
 Author 4: Sushank Gopala

International Journal of Advanced Computer Science and Applications(IJACSA), Volume 15 Issue 6, 2024.

[Abstract and Keywords](#)[How to Cite this Article](#)[BibTeX Source](#)

**Abstract:** Deep learning has revolutionized industries such as natural language processing and computer vision. This study explores the fusion of these domains by proposing a novel approach for text extraction and translation using lip reading and deep learning. Lip reading, the process of interpreting spoken language by analyzing lip movements, has garnered interest due to its potential applications in noisy environments, silent communication, and accessibility enhancements. This study employs the power of deep learning architectures such as CNNs and RNNs to accurately extract text content from lip movements captured in video sequences. The proposed model consists of multiple stages: lip region detection, feature extraction, text recognition, and translation. Initially, the model identifies and isolates the lip region within video frames using a CNN-based object detection approach. Subsequently, relevant features are extracted from the lip region using CNNs to capture intricate motion patterns and convert these visual features into textual information. The extracted text is further processed and translated into the desired language using machine translation techniques to enable translation.

 **Keywords:** Deep Learning (DL); Convolutional Neural Networks (CNN); Lip Reading; Recurrent Neural Networks (RNN)



### Upcoming Conferences



**Future of Information and Communication Conference (FICC) 2025**

28-29 April 2025

 Berlin, Germany



**Computing Conference 2025**

19-20 June 2025



Advertisement

WILEY

我们的全新语言检查工具可在**一分钟内**完成文章语言语法的评估

## Expert Systems



ORIGINAL ARTICLE

### Energy optimization in path arbitrary wireless sensor network

B. Harish Goud T. N. Shankar, Basant Sah, Rajanikanth Aluvalu

First published: 16 March 2023 | <https://doi.org/10.1111/exsy.13282> | Citations: 2[Read the full text >](#)

PDF



TOOLS



SHARE

Advertisement



Volume 41, Issue 2

Special Issue: Machine Learning Challenges and Applications for Industry 4.0 (EXSYS-MLI4.0) / Intelligent Computational Methods for Economics

February 2024

e13282

### Abstract

A network of wireless sensors is a self-infrastructure approach with many sensory nodes. The distributed sensory nodes communicate with each other via sensory points. In wireless sensor network (WSN), the sensory nodes collect information for healthcare, military and monitoring systems. Such networks require an exclusive arrangement of the nodes to challenge inherent limitations and energy deficiency. The conventional design of a communication system consumes more energy with high latency causing degraded performance. This study provided a machine learning based path optimization

nature > scientific reports > articles > article

Article | Open access | Published: 08 January 2024

# An improved GBSO-TAENN-based EEG signal classification model for epileptic seizure detection

M. V. V. Prasad Kantipudi, N. S. Pradeep Kumar, Rajanikanth Aluvalu, Shitharth Selvarajan & K Kotecha

Scientific Reports 14, Article number: 843 (2024) | Cite this article

2565 Accesses | 6 Citations | Metrics

## Abstract

Detection and classification of epileptic seizures from the EEG signals have gained significant attention in recent decades. Among other signals, EEG signals are extensively used by medical experts for diagnosing purposes. So, most of the existing research works developed

Download PDF



Sections

Figures

References

Abstract

Introduction

Related works

GBSO-TAENN-based seizure prediction

Results and discussion

Conclusion

Data availability

References

SEARCH ▾

BROWSE ▾

Recent Searches

My List

My PsycNet

## RETRACTED ARTICLE: Designing a cognitive smart healthcare framework for seizure prediction using multimodal convolutional neural network.

[EXPORT](#) [★ Add To My List](#) [✉](#) [🖨](#) [🔗](#) [🔗](#)

Database: APA PsycInfo

Retraction

### Citation

[Full text from publisher](#)

Aluvalu, R., Aravinda, K., Maheswari, V. U., Kumar, K. A. J., Rao, B. V., & Prasad, K. M. V. V. (2024). RETRACTED ARTICLE: Designing a cognitive smart healthcare framework for seizure prediction using multimodal convolutional neural network. *Cognitive Neurodynamics*, 18(6), 4107. <https://doi.org/10.1007/s11571-023-10049-x>

### Abstract

Reports the retraction of "Designing a cognitive smart healthcare framework for seizure prediction using multimodal convolutional neural network" by Rajanikanth Aluvalu, K. Aravinda, V. Uma Maheswari, K. A. Jayasheel Kumar, B. Venkateswara Rao and Kantipudi M. V. V. Prasad (*Cognitive Neurodynamics*, Advanced Online Publication, Jan 2, 2024, np). The Editor-in-Chief and the publisher have retracted this article. The article was submitted to be part of a guest-edited issue. An investigation by the publisher found a number of articles, including this one, with a number of concerns, including but not limited to compromised editorial handling and peer review process, inappropriate or irrelevant references or not being in scope of the journal or guest-edited issue. Based on the investigation's findings the Editor-in-Chief therefore no longer has confidence in the results and conclusions of this article. The authors have not responded to correspondence from the publisher regarding this retraction. The online version of this article contains the full text of the retracted article as Supplementary Information. (The abstract of the original article appears below.) The Internet of Things (IoT) has evolved from a network of embedded computer devices to a network of brainy sensors. With the advent of IoT-cloud technologies, however, there is a growing demand for a cognitive framework that can deliver affordable, high-quality smart healthcare with a focus on the discrete patient. The advent of AI and deep learning methods makes it possible to include human-level reasoning into intelligent healthcare infrastructure. To evaluate the healthcare system, we also

 View PDF

Download full issue

## Outline

Abstract

Keywords

1. Introduction

2. Related work

3. Failure prediction model

4. Result analysis

5. Discussion

6. Conclusions and future work

Declaration of competing interest

References

Show full outline 

Cited by (10)

Figures (17)






## High-Confidence Computing


Volume 3, Issue 4, December 2023, 100165



Research article

# Machine learning job failure analysis and prediction model for the cloud environment

 Harikrishna Bommala <sup>a</sup>, Uma Maheswari V. <sup>b</sup>, Rajanikanth Aluvalu <sup>c</sup>  , Swapna Mudrakola <sup>d</sup>
Show more 
 Add to Mendeley  Share  Cite

<https://doi.org/10.1016/j.hcc.2023.100165>
[Get rights and content](#)
[Under a Creative Commons license](#)
 open access

## Abstract

Reliable and accessible cloud applications are essential for the future of ubiquitous computing, smart appliances, and electronic health. Owing to the vastness and diversity of the cloud, a most cloud services, both physical and logical services have failed. Using

Recommended articles 

## An energy-efficient resource allocation strategy in massive MIMO-enabled...

 High-Confidence Computing, Volume 3, Issue 3, 2023, ...  
 Yibin Xie, ..., Yang Zhang

 View PDF

## Fault-tolerant scheduling of graph-based loads on fog/cloud environments with...

 Computer Networks, Volume 235, 2023, Article 109964  
 Felor Beikzadeh Abbasi, ..., Ali Movaghar

 View PDF

## Reaching consensus for membership dynamic in secret sharing and its...

 High-Confidence Computing, Volume 3, Issue 3, 2023, ...  
 Yan Zhu, ..., Haibin Zheng

 View PDF

[Show 3 more articles](#) 

Article Metrics

 FEEDBACK

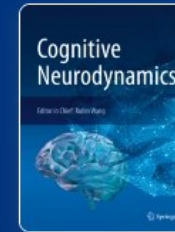
Home > [Cognitive Neurodynamics](#) > Article

# RETRACTED ARTICLE: Designing a cognitive smart healthcare framework for seizure prediction using multimodal convolutional neural network

Research Article | Published: 02 January 2024

Volume 18, page 4107, (2024) [Cite this article](#)


Download PDF 





## Cognitive Neurodynamics

[Aims and scope](#) →

[Submit manuscript](#) →

[Rajanikanth Aluvalu](#) , [K. Aravinda](#), [V. Uma Maheswari](#), [K. A. Jayasheel Kumar](#), [B. Venkateswara Rao](#) & [Kantipudi M. V. V. Prasad](#)

 664 Accesses  4 Citations [Explore all metrics](#) →

 This article has been [updated](#)

[Use our pre-submission checklist](#) →

Avoid common mistakes on your manuscript.



### Sections

[Change history](#)





Home → International Journal of Vehicle Autonomous Systems → Vol. 16, No. 2-4

NO ACCESS

# Blockchain and IoT architectures in autonomous vehicles

Rajanikanth Aluvalu, V. Uma Maheswari, Swapna Mudrakola and Krishna Keerthi Chennam

Published Online: August 22, 2023 · pp 180-203 · <https://doi.org/10.1504/IJVAS.2022.133010>



PDF



Tools

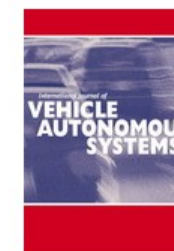


Share

## Abstract

In this new digital era, the automotive industry integrates various technologies like 5G, AI, IoT and blockchain, etc., to enhance communication and security. The automotive industry is making efforts to develop fully autonomous vehicles. This article discusses the opportunities, challenges and limitations of integrating blockchain and IoT for developing autonomous vehicles. Various types of autonomous vehicles and security concerns in autonomous vehicles are discussed in this article. IoT helps develop smarter vehicles through In-vehicle communication, Vehicle-to-Vehicle and Vehicle to infrastructure communication. Security is the major concern with IoT. Integrating blockchain technology with IoT-enabled vehicles will enhance data security. The blockchain uses cryptographic hashes to record and store data, which will help protect the data from unauthorised access. Retrospectively, the article provides the comprehensive study of integrating IoT and blockchain technologies for

Figures References Related Details



Volume 16 · Issue 2-4 · 2022

ISSN: 1471-0226  
eISSN: 1741-5306

### History

Published Online: August 22, 2023

Copyright © 2022 Inderscience Enterprises Ltd.

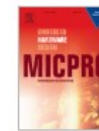
[View PDF](#)[Download full issue](#)

## Outline

[Abstract](#)[Keywords](#)[1. Introduction](#)[2. Related work](#)[3. Digital twins technology in healthcare system](#)[4. Novel proposed approach for emergency hosp...](#)[5. Methodology](#)[6. Advantages of proposed model and limitations](#)[7. Results & discussion](#)[8. Conclusion & future scope](#)[Funding](#)[Declaration of Competing Interest](#)[Data availability](#)[References](#)[Vitae](#)

## Microprocessors and Microsystems

Volume 98, April 2023, 104794



## The novel emergency hospital services for patients using digital twins

Rajanikanth Aluvalu <sup>a</sup>, Swapna Mudrakola <sup>b</sup>, Uma Maheswari V <sup>c</sup>, A.C. Kaladevi <sup>d</sup>, M.V.S Sandhya <sup>e</sup>, C. Rohith Bhat <sup>f</sup>[Show more](#)[+ Add to Mendeley](#) [Share](#) [Cite](#)<https://doi.org/10.1016/j.micpro.2023.104794>[Get rights and content](#)

## Abstract

The Digital twins will duplicate the actual objects, create the virtual world and execute using IoT devices and Sensors. The Emergency Room Service (ERS) is a critical phase for patients in health condition evaluation, Digital Health records will help us in understanding the cause of illness, medical history will help us to start the treatment.

Part of special issue

[Enabling Technologies for Intelligent Embedded Systems: Changing the Landscape of Research and Development](#)

Edited by Muhammad Shafique, V.Vinoth Kumar, Ahmed A. Elgar, Polinpapilinho F. Katina

[View special issue](#)

Recommended articles

[Intelligent Digital Twin in Health Sector: Realization of a Software-Service for...](#)IFAC-PapersOnLine, Volume 55, Issue 19, 2022, pp. 79-...  
Samira Maleki, ..., Behrang Ashtari[View PDF](#)[Using Digital Twins for Precision Medicine in Vascular Surgery](#)Annals of Vascular Surgery, Volume 67, 2020, pp. ...  
Fabien Lavra, Juliette Raffort[FEEDBACK](#)



1 of 1

Download Print Save to PDF Add to List Create bibliography

 Page could not be loaded  
Please reload to try again

[Reload page](#)

Cited by 0 documents

Inform me when this document is cited in Scopus:

[Set citation alert >](#)

Related documents

[Energy-efficient and secure routing strategy for opportunistic data transmission in WSNs](#)

Narayana, P. , Keerthi, K. , Khalaf, O.I. (2024) *Journal of Cyber Security Technology*

[An enhanced bio-inspired energy-efficient localization routing for mobile wireless sensor network](#)

Tumula, S. , Rama Devi, N. , Ramadevi, Y. (2024) *International Journal of Communication Systems*

References (47)

[View in search results format >](#)


All Export Print E-mail Save to PDF Create bibliography

1 Meenakshi, B., Karunkuzhali, D.  
[Enhancing cyber security in WSN using optimized self-attention-based](#)  
[positional-unintentional auto-encoder generative adversarial network](#)



1 of 1

Download Print Save to PDF Add to List Create bibliography

 Page could not be loaded  
Please reload to try again

[Reload page](#)

Cited by 1 document

A New Metaheuristic Approach to Diagnosis of Parkinson's Disease Through Audio Signals

Oguz, O. , Badem, H. (2024) *Elektronika ir Elektrotechnika*

[View details of this citation](#)

Inform me when this document is cited in Scopus:

[Set citation alert >](#)

References (12)

[View in search results format >](#)

All Export Print E-mail Save to PDF Create bibliography

1 Moro-Velazquez, L., Gomez-Garcia, J.A., Arias-Londoño, J.D., Dehak, N., Godino-Llorente, J.I.  
[Advances in Parkinson's Disease detection and assessment using voice and speech analysis of the articulators and laboratory aspects](#)

Related documents

Diagnosis of Parkinson's Disease Using Deep Neural Network Model

Anila, M. , Pradeepini, G. (2021) *2021 International Conference on*



1 of 1

Download Print Save to PDF Add to List Create bibliography

 Page could not be loaded  
Please reload to try again  
[Reload page](#)

References (23)

[View in search results format >](#)

All Export Print E-mail Save to PDF Create bibliography

1 Ahmad, R., Wazirali, R., Abu-Ain, T.  
[Machine Learning for Wireless Sensor Networks Security: An Overview of](#)

Cited by 3 documents

Multi-view consistent generative adversarial network for enhancing intrusion detection with prevention systems in mobile ad hoc networks against security attacks

Rajkumar, M. , Karthika, J. , Abinayaa, S.S. (2025) *Computers and Security*

Enhancing intrusion detection in MANETs with blockchain-based trust management and enhanced GRU model

Krishna, E.S.P. , Sandeep, D. , Kocherla, R. (2025) *Peer-to-Peer Networking and Applications*

A secure routing and black hole attack detection system using coot Chimp Optimization Algorithm-based Deep Q Network in MANET

D, S. , PH, L. (2025) *Computers and Security*

[View PDF](#)[Download full issue](#)

## Outline

[Abstract](#)[Keywords](#)[1. Introduction](#)[2. Related work](#)[3. Proposed FPISMF model](#)[4. Experimental evaluation](#)[5. Conclusion and future work](#)[Declaration of competing interest](#)[Data availability](#)[References](#)[Show full outline](#)[Cited by \(1\)](#)[Figures \(6\)](#)

## Measurement: Sensors

Volume 33, June 2024, 101164



# Fake product identification for small and medium firms (FPISMF) using blockchain technology

Sangeeta Gupta <sup>a</sup> , Ramu Kuchipudi <sup>b</sup> , Md Sohail <sup>a</sup> , Karan Singh <sup>a</sup> , J. Mahalakshmi <sup>c</sup> , Ashok Sarabu <sup>d</sup> [Show more](#)[+ Add to Mendeley](#) [Share](#) [Cite](#)<https://doi.org/10.1016/j.measen.2024.101164>[Get rights and content](#)Under a Creative Commons license [↗](#)

open access

## Abstract

Counterfeit products have become a significant problem for small and medium-sized businesses (SMBs) with the estimated value of counterfeit goods worldwide reaching

Part of special issue [^](#)

## Device Condition Monitoring and Predictions using IoT in Industry 5.0

Edited by Praveen Kumar Donta, Yu-Chen Hu, Abhishek Hazra, Ihsan Ali

[View special issue](#)Recommended articles [^](#)

## A novel approach for solar combined open-ended winding induction machine for...

Measurement: Sensors, Volume 33, 2024, Article 101141  
V. Kavya, ..., T. Jarin[View PDF](#)

## Design and implementation of a monitoring platform based on beidou high precision...

Measurement: Sensors, Volume 33, 2024, Article 101105  
Xiusheng Ren, ..., Qiang Liang[View PDF](#)[FEEDBACK](#)

All

ADVANCED SEARCH

Journals & Magazines > IEEE Access > Volume: 12

# Empowering Cybersecurity Using Enhanced Rat Swarm Optimization With Deep Stack-Based Ensemble Learning Approach

Publisher: IEEE

[Cite This](#)

[PDF](#)

P. Manickam ; M. Girija ; Ashit Kumar Dutta ; Palamakula Ramesh Babu ; Krishan Arora ; Mun Jeong [All Authors](#)

2  
Cites in  
Papers

594  
Full  
Text Views



Open Access Comment(s)

Under a Creative Commons License

## Abstract

### Abstract:

Cybersecurity is a vital technology and measures intended to protect networks, computers, information, and

Setting math: 30%  
Document Sections

Try out IEEE's Experimental

## AI-Based Search

[Try it Now!](#)

### More Like This

- Hyperparameter Optimization
- Long Short Term Memory Models for Interpretable Electrical Fault Classification

IEEE Access

[PDF](#)

[Help](#)

[Feedback](#)

[nature](#) > [scientific reports](#) > [articles](#) > [article](#)Article | [Open access](#) | Published: 16 September 2023

# Prediction of DDoS attacks in agriculture 4.0 with the help of prairie dog optimization algorithm with IDNet

[Ramesh Vatambeti](#) , [D. Venkatesh](#), [Gowtham Mamidiseti](#), [Vijay Kumar Damera](#), [M. Manohar](#) & [N. Sudhakar Yadav](#)[Scientific Reports](#) **13**, Article number: 15371 (2023) | [Cite this article](#)1886 Accesses | 4 Citations | 4 Altmetric | [Metrics](#)

## Abstract

Integrating cutting-edge technology with conventional farming practices has been dubbed “smart agriculture” or “the agricultural internet of things.” Agriculture 4.0, made possible by

Download PDF



Sections

Figures

References

[Abstract](#)[Introduction](#)[Related works](#)[Proposed system](#)[Results and discussion](#)[Conclusions](#)[Data availability](#)[References](#)



## Performance Analysis of Sparks Machine Learning Library

Seyedfaraz Yasrobi, Jakayla Alston, Babak Yadranjiaghdam, Nasschzadeh Tabrizi

East Carolina University, Department of Computer Science  
Science and Technology Building  
Greenville, NC 27858-4353 USA

yasrobis14@students.ecu.edu, alstonja14@students.ecu.edu, yadranjiagh-  
damb15@studentds.ecu.edu, tabrizim@ecu.edu

**Abstract.** This paper examines the performance of Apache Sparks machine learning library with reference to the optimal required resources such as the number of machines and cores to best perform popular machine learning algorithms. In order to achieve this, we have observed the training time of classification algorithms such as logistic regression, support vector machines, decision trees, random forests, and gradient boosted trees under different configurations on a sample dataset. Our research revealed that having an excessive number of resources does not necessarily decrease the training time of the machine learning algorithms, rather, it may even degrade the training time by up to 30 percent. Furthermore, this study confirms that methodologies such as tree ensembles can increase the training time of machine learning algorithms compared to that of typical decision trees.

### 1 Introduction

Machine learning with its automated learning power unleashes big data power to aid data scientists to gain knowledge in a variety of applications such as computer vision, speech processing, natural language understanding, neuroscience, health, and Internet



DOI: 10.1080/01611194.2022.2109943 • Corpus ID: 252825807

Share

# Cryptanalysis of RSA with small difference of primes and two decryption exponents: Jochemsz and May approach

R. S. Kumar, R. M. K. Sureddi • Published in *Cryptologia* 10 October 2022 • Computer Science, Mathematics

**TLDR** This attack examines the closeness of the primes chosen whenever the RSA system is used for two instances with the same modulus, and the bound is highly efficient compared to other known attacks. [Expand](#)

3 Citations

[View All](#)

- [View on Taylor & Francis](#)
- [Save to Library](#)
- [Create Alert](#)
- [Cite](#)

Topics

3 Citations

50 References

Related Papers

## Topics

AI-Generated

Modulus

Rivest-Shamir-Adleman

Decryption Exponent

Cryptanalysis

LLL Algorithm

Cryptosystem

By clicking accept or continuing to use the site, you agree to the terms outlined in our [Privacy Policy](#), [Terms of Service](#), and [Dataset License](#)

ACCEPT & CONTINUE

Home > [Journal of Computer Virology and Hacking Techniques](#) > Article

# Partial key exposure attack on RSA using some private key blocks

Original Paper | Published: 08 November 2023

Volume 20, pages 185–193, (2024) [Cite this article](#)



## Journal of Computer Virology and Hacking Techniques

[Aims and scope](#) →

[Submit manuscript](#) →

Download PDF ↓

✓ Access provided by CBIT–Library & Information Centre Hyderabad

[Santosh Kumar Ravva](#) ✉, [K. L. N. C. Prakash](#) & [S. R. M. Krishna](#)

📄 298 Accesses [Explore all metrics](#) →

## Abstract

[Use our pre-submission checklist](#) →

Avoid common mistakes on your manuscript.

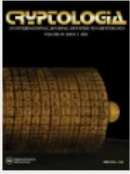


Sections

Figures

References

Abstract



Cryptologia >

Volume 49, 2025 - Issue 1

Submit an article

Journal homepage

Enter keywords, authors, DOI, etc

This Journal



Advanced search

86

Views

1

CrossRef citations to date

0

Altmetric

Articles

# An improved cryptanalysis of multi-prime RSA with specific forms of decryption exponent

Santosh Kumar R, Prakash Klnc & Krishna Srm

Pages 1-14 | Published online: 05 Jan 2024

Cite this article <https://doi.org/10.1080/01611194.2023.2271463>



Full Article

Figures & data

References

Citations

Metrics

Reprints & Permissions

Read this article

Share

## Abstract

Multi-prime RSA (MPRSA) is an extended version of RSA in which the modulus is the product of three or more distinct prime numbers. As with the RSA, this variant is also scrutinized for vulnerabilities. The main goal of the attacks on MPRSA is to test the strength of MPRSA compared to that of RSA to determine whether MPRSA can be used in place of RSA. We apply two famous attacks on RSA to MPRSA

## Related Research

People also read

Recommended articles

Cited by 1

Cryptanalysis of RSA with small difference of primes and two decryption exponents: Jochemsz and May approach

Sample our Computer Science Journals >> Sign in here to start your access to the latest two volumes for 14 days


Home > Journal of Computer Virology and Hacking Techniques > Article

# Cryptanalysis of RSA with composed decryption exponent with few most significant bits of one of the primes

Original Paper | Published: 20 October 2023

Volume 20, pages 195–202, (2024) [Cite this article](#)

Download PDF 

 Access provided by CBIT–Library & Information Centre Hyderabad



## Journal of Computer Virology and Hacking Techniques

[Aims and scope](#) →

[Submit manuscript](#) →

[R. Santosh Kumar](#) , [K. L. N. C. Prakash](#) & [S. R. M. Krishna](#)

 265 Accesses  2 Citations [Explore all metrics](#) →

## Abstract

RSA is well known public-key cryptosystem in modern-day cryptography. Since the

[Use our pre-submission checklist](#) →

Avoid common mistakes on your manuscript.



Sections

References

[Abstract](#)

Home > Full-text access for editors

### **[Cryptanalysis of common prime RSA with two decryption exponents: Jochemsz and May approach](#)**

by Santosh Kumar Ravva; Sureddi R.M. Krishna

*International Journal of Information and Computer Security (IJICS), Vol. 22, No. 3/4, 2023*

**Abstract:** RSA is a well-known public key cryptosystem in modern-day cryptography. Common prime RSA (CP-RSA) is a variant of RSA which is introduced by Wiener to avoid the small secret exponent attack on RSA. Lattice-based reduction algorithms were successfully used for cryptanalysis for RSA and its variants. In this paper, we mount an attack on CP-RSA by following the Jochemsz and May approach. Jochemsz and May approach is the standard way to construct the lattices for the attacks on RSA and its variants. Our attack improves the bounds of attacks on standard RSA and CP-RSA.

*Online publication date: Tue, 09-Jan-2024*

The full text of this article is only available to individual subscribers or to users at subscribing institutions.

#### **Existing subscribers:**

Go to [Inderscience Online Journals](#) to access the [Full Text](#) of this article.

#### **Pay per view:**

If you are not a subscriber and you just want to read the full contents of this article, [buy online access here](#).


#### **Complimentary Subscribers, Editors or Members of the Editorial Board of the International Journal of Information and Computer Security (IJICS):**

Login with your Inderscience username and password:

#### Keep up-to-date

 [Our Blog](#)

 [Follow us on Twitter](#)

 [Visit us on Facebook](#)

 [Our Newsletter \(subscribe for free\)](#)

 [RSS Feeds](#)

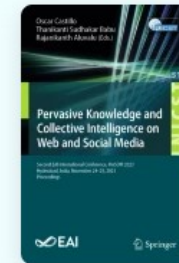
 [New issue alerts](#)

Home > [Pervasive Knowledge and Collective Intelligence on Web and Social Media](#) > Conference paper

# Citation Analysis of a Q1 Journal from Its Thirty Years of Inception in Agriculture Supply-Chain

Conference paper | First Online: 13 August 2024

pp 101–109 | [Cite this conference paper](#)



**Pervasive Knowledge and Collective Intelligence on Web and Social Media**  
(PerSOM 2023)

[Pragati Priyadarshinee](#) & [M. V. V. Prasad Kantipudi](#)

Part of the book series: [Lecture Notes of the Institute for Computer Sciences, Social Informatics](#)

Access this chapter

[Login via an institution](#)



Advertisement

WILEY

Bring your research to  
new audiences

Promote your research using a Graphical  
Abstract designed by Wiley Editing Services!



## SECURITY AND PRIVACY

RESEARCH ARTICLE

### Anomaly detection in IoT environment using machine learning

Harini Bilakanti, Sreevani Pasam, Varshini Palakollu, Sairam Utukuru

First published: 01 January 2024 | <https://doi.org/10.1002/spy2.366> | Citations: 1

[Read the full text >](#)



PDF



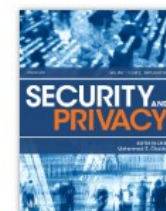
TOOLS



SHARE

#### Abstract

This research paper delves into the security concerns within Internet of Things (IoT) networks, emphasizing the need to safeguard the extensive data generated by interconnected physical devices. The presence of anomalies and faults in the sensors and devices deployed within IoT networks can significantly impact the functionality and outcomes of IoT systems. The primary focus of this study is the identification of anomalies in IoT devices arising sensor tampering, with an emphasis on the application of machine learning techniques. While supervised methods like one-class SVM, Gaussian



Volume 7, Issue 3  
May/June 2024  
e366

Advertisement

WILEY

#### Internet Technology Letters

##### Call for Papers

Special Issue:  
TinyML for Edge  
Intelligence and its  
Applications in IoTs

Deadline:  
30 April 2025





# A MACHINE LEARNING APPROACH IN COMMUNICATION 5G-6G NETWORK

SWATI LAKSHMI BOPPANA<sup>1</sup>, ANNAPURNA GUMMADI<sup>2</sup>, YALLAPRAGADA RAVI RAJU<sup>3</sup>,  
DR. SATISH THATAVARTI<sup>4</sup>, RAMU KUCHIPUDI<sup>5</sup>, TENALI ANUSHA<sup>6</sup>, RALLABANDI CH S N  
P SAIRAM<sup>7</sup>

<sup>1</sup>Department of ECE, Prasad V. Potluri Siddhartha Institute of Technology, Vijayawada, India

<sup>2</sup>Department of CSE (Data Science), CVR College of Engineering, Hyderabad, India.

<sup>3</sup>Department of CS&T, Madanapalle Institute of Technology & Science, Madanapalle, India

<sup>4</sup>Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India

<sup>5</sup>Department of IT, Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, India

<sup>6</sup>Department of AIML, School of Engineering, Mallareddy University, Hyderabad, India

<sup>7</sup>Department of CSE (Data Science), R.V.R. & J. C. College of Engineering, Guntur, India

boppanaswathi@gmail.com, gummadiannapurna@gmail.com, ravirajuy@mits.ac.in,  
drsatishtathavarti@kluniversity.in, kramupro@gmail.com, anusharajburuga@gmail.com,  
sairam.mtech20@gmail.com

## ABSTRACT

Applications in the fields of entertainment, commerce, health, and public safety rely heavily on wireless communication technologies. These technologies are constantly improving with each new generation, and the latest example of this is the widespread implementation of 5G wireless networks. Industry and academia are already planning the next generation of wireless technologies, 6G, which will be an improvement over 5G. When it comes to 6G systems, one of the most important things is that these wireless networks employ AI and ML. There will be some kind of artificial intelligence or machine learning used in every part of a wireless system that we know about from our experience with wireless technologies up to 5G, including the physical, network, and application levels. A current overview of concepts for future wireless networks, including 6G, and the relevance of ML approaches in these systems is presented in this overview article. Specifically, we set out a 6G conceptual model and demonstrate how ML approaches are utilized and contribute to each layer of the model. With wireless communication systems in mind, we take a look back at many ML methods, both old and new, including supervised and unsupervised learning, RL, DL, and FL.

DOI: [10.14569/IJACSA.2024.0150253](https://doi.org/10.14569/IJACSA.2024.0150253)

## Deep Learning Augmented with SMOTE for Timely Alzheimer's Disease Detection in MRI Images

[PDF](#)

Author 1: P Gayathri Author 2: N. Geetha Author 3: M. Sridhar Author 4: Ramu Kuchipudi  
Author 5: K. Suresh Babu Author 6: Lakshmana Phaneendra Maguluri Author 7: B Kiran Bala

International Journal of Advanced Computer Science and Applications(IJACSA), Volume 15 Issue 2, 2024.

[Abstract and Keywords](#)[How to Cite this Article](#)[BibTeX Source](#)

**Abstract:** Timely diagnosis of Alzheimer's Disease (AD) is pivotal for effective intervention and improved patient outcomes, utilizing Magnetic Resonance Imaging (MRI) to unveil structural brain changes associated with the disorder. This research presents an integrated methodology for early detection of Alzheimer's Disease from Magnetic Resonance Imaging, combining advanced techniques. The framework initiates with Convolutional Neural Networks (CNNs) for intricate feature extraction from structural MRI data indicative of Alzheimer's Disease. To address class imbalance in medical datasets, Synthetic Minority Over-sampling Technique (SMOTE) ensures a balanced representation of Alzheimer's Disease and non- Alzheimer's Disease instances. The classification phase employs Spider Monkey Optimization (SMO) to optimize model parameters, enhancing precision and sensitivity in Alzheimer's Disease diagnosis. This work aims to provide a comprehensive approach, improving accuracy and tackling imbalanced datasets challenges in early Alzheimer's detection. Experimental outcomes demonstrate the proposed approach outperforming conventional techniques in terms of classification accuracy, sensitivity, and specificity. With a notable 91% classification accuracy, particularly significant in medical diagnostics, this method holds promise for practical application in clinical settings, showcasing robustness and potential for enhancing patient outcomes in early-stage Alzheimer's diagnosis. The implementation is conducted in Python.



### Upcoming Conferences



**Future of Information and Communication Conference (FICC) 2025**

28-29 April 2025

[Berlin, Germany](#)



**Computing Conference 2025**

19-20 June 2025

# AI ENABLED SYSTEM WITH REAL TIME MONITORING OF PUBLIC SURVEILLANCE VIDEOS FOR ABNORMALITY DETECTION AND NOTIFICATION

PRATHAP ABBAREDDY<sup>1</sup> DR. SK. YAKOOB<sup>2</sup> RAMU KUCHIPUDI<sup>3</sup> BHUJANGA REDDY BHAVANAM<sup>4</sup>

<sup>1</sup>Software Engineer

<sup>2</sup>Associate Professor &HOD, Dept. of CSE, Sai Spurthi Institute of Technology  
B.Gangaram,Sathupally, Khammam-507303 ,Telangana,

<sup>3</sup>Associate Professor, Chaitanya Bharathi Institute of Technology, Department  
of Information Technology, Gandipet, Hyderabad, Telangana -500075

<sup>4</sup>Asst.Professor computer science and engineering, Geethanjali college of engineering and technology

Email :prathap.abbareddy@gmail.com <sup>2</sup>yakoobcs2004@gmail.com <sup>3</sup>kramupro@gmail.com <sup>4</sup>

<sup>4</sup>bhujangareddy.cse@gcet.edu.in

## ABSTRACT

Public surveillance videos are increasingly playing key role in identification of certain incidents and people who misbehave or perform illegal activities. Monitoring surveillance videos manually to detect abnormalities is time consuming and it may lead to delay in getting required information. With the usage of Artificial Intelligence (AI) video analytics in real time can help in acquiring such information on time so as to make well informed decisions. Particularly deep learning is great help in learning from incidents and detect anomalous behaviours. In this study, we suggested an autonomous system for anomaly detection from surveillance films, based on deep learning. For anomaly detection, an improved Convolutional Neural Network (CNN) model is employed. We presented a method that utilizes the upgraded CNN model for its functionality, called Learning based Video Anomaly Detection (LbVAD). To lower the prediction process's error rate, a loss function is defined. For our empirical investigation, we gathered data from many benchmark datasets, including UMN, UCSD, Ped1, and Ped2. The suggested approach works better than the current models, according to the results of our experiments.

**Keywords:** Machine Learning, Deep Learning, Artificial Intelligence, Video Abnormality Detection

## 1. INTRODUCTION

# Identification of Lung Cancer Using Ensemble Methods Based on Gene Expression Data

K. Mary Sudha Rani<sup>1\*</sup>, Dr. V. Kamakshi Prasad<sup>2</sup>

Submitted: 28/05/2023

Revised: 06/07/2023

Accepted: 25/07/2023

**Abstract:** Lung cancer is consistently classified as the most dangerous form of the disease since the beginning of recorded history. Patients with lung cancer who receive appropriate medical care, such as a low-dose CT scan, have a far better chance of survival since the disease is detected and diagnosed early. Nonetheless, there are certain drawbacks to this attempt. The gene expression level in hundreds of genes or cells within each tissue may now be determined because of developments in DNA microarray technology. Even though machine learning (ML) is rapidly being used in the medical field for lung cancer detection, the shortage of interpretability of these models remains a significant hurdle. Machine learning can be used to analyze gene expression data (DNA microarray) to predict whether or not a patient has lung cancer. The Collective Random Forest and Adaptive Boosting were employed to determine who was responsible for the harm. KPCA, or Kernel principal component analysis, was used for the feature reduction procedure. We calculated the correlation between each feature and the target using the statistical parameters provided by KPCA. Determining the proportion of the correct predictions for a given data set is one way to calculate the accuracy of a classification model. We tested the validity of the proposed technique in this work using a dataset including information about lung cancer. The dataset includes GSE4115 from the Gene Expression Omnibus (GEO) database, as well as the expression profiles it contains. The findings demonstrate the Identification of Lung Cancer (IOLC) model's potential to detect lung cancer in terms of accuracy, precision, recall, F-Measure, and error rate, with results indicating an accuracy of 81%, the precision of 81.2%, recall of 78.9%, F-Measure of 77.7%, and error rate of 0.29%, respectively.

**Keywords:** Gene Expression, Lung cancer, Ensemble machine learning Random forest, AdaBoost

## 1. Introduction

Cancer is a disease that causes cell destruction in the body. Cells develop and increase in a controlled manner; nevertheless, this control may fail if an error occurs in the cell's genetic blueprints. A variety of factors can cause this mistake. Lung cancer is the most common and lethal malignant tumor seen worldwide. In 2012, around 1,800,000 new lung cancer cases were detected, with 1,600,000 people dying due to the condition. Lung cancer is more common in women and is the leading cause of cancer death. Although smoking is the primary cause of lung cancer, around 15% of male and 53% of female lung cancer patients did not smoke. Furthermore, it is estimated that 25% of lung cancer patients worldwide did not get the disease due to smoking. Previously, the primary resource in biology was gene networks GNs [1], commonly depicted as graphs with nodes and rods, with nodes representing genes and rods signifying gene interactions. These rods may be assigned a numerical number or weight based on the strength of the relationships between the parties involved. As a result, GNs can uncover genes linked with biological processes and their interactions, providing a complete picture of the processes under inquiry. GNs are widely utilized in many

fields, inquiry. GNs are widely utilized in many fields, including but not limited to biology, healthcare, and bioinformatics.

Furthermore, when it comes to non-smokers have a different carcinogenic pathway, clinic pathological features, epidemiology, and natural history than smokers. Lung cancer is the most common type diagnosed worldwide and the leading cause of cancer fatalities wheezing, hoarseness, chest tightness, coughing, and spitting up blood are all indications of lung cancer. Indications and symptoms include chest discomfort, shortness of breath, and wheezing [2]. To avoid this dreadful situation, we require machine learning algorithms to aid in the early detection and prevention of lung cancer. Treatments can be more effective and less likely to recur if started early in lung cancer [3]. As a result, preventative lung cancer screening and detection may be therapeutically beneficial, particularly for patients with undiagnosed lung disease. Experiments have uncovered the genes responsible for lung cancer mutagenicity and pathogenesis, albeit most genes have only a tenuous link to the disease. To determine whether a gene is linked to lung cancer, one must run several trials, which would necessitate a considerable financial commitment.

On the other hand, machine learning (ML) algorithms can prioritize disease-triggering genes where their significant

<sup>1</sup>\*Research scholar, Dept. of CSE, JNTUH, Assistant Professor, CSE Dept., Chaitanya Bharathi of Technology Hyderabad, Telangana, India  
Email: kmarysudha\_cse@cbt.ac.in

<sup>2</sup>Professor, Dept. of CSE, JNTUH, Telangana, India

studies offer the ability to uncover the association amid cancer identification genes. These findings can potentially be used in the early detection of cancer. In particular, successful ML approaches are described works. These algorithms include artificial neural network-based computer-aided diagnosis [4, 5], ensemble approaches [6, 7], and hybrid methods.

Because of its extensive prevalence, it has been examined for cancer biomarkers that can predict a disease's prognosis. To be more exact, lung carcinoma is one of the most common types of cancer, with tobacco use accounting for over 85 percent of cases. Regrettably, the vast majority of instances are fatal. This is due, in part, to a delayed diagnosis, which necessitates specialized medical procedures such as bronchoscopy. As a result, lung cancer biomarkers are seen as critical in the disease's early identification; as a response, numerous initiatives have investigated non-invasive approaches for testing these biomarkers [8]. Ensemble learning could be effective in our investigation because it can increase a model's robustness and accuracy by merging multiple imperfect classifiers. Ensemble learning, which includes bagging and boosting, can be viewed as a general bagging technique for enhancing cancer classification. Random forest is another popular bagging technique. Gene microarray (GMA) recently appeared as a promising cancer detection and classification technique. Statistical analysis and machine learning methods were used to uncover accurate gene characteristics that can be used as inputs for cancer classification models. The lung cancer data's limited sample size makes the interpretation and training of microarray data problematic. The presence of noise in the samples can have a negative impact on the training models' performance. Furthermore, the random forest was utilized to search the classifiers at random, and in the training stage, a better judgment was generated.

One technique like self-paced learning, while another develops a novel formulation to uncover samples [10]. As the SPL regularizer's penalty steadily increases during optimization, more samples during the training phase are selected modes. Adoption has been quick, especially in multi-task learning [11], image categorization [12], and molecular descriptor selection [13, 14]. So, in the current article, we extract high-quality samples using this strategy. The following contributions were made by this paper, which is given below.

- We initially proposed using the IOLC to train a cancer detection and classification model using DNA microarray technology. The samples' degrees of confidence ranged from high to low.
- We built a machine learning prediction model using the Ensemble Methods Random Forest, and AdaBoost was the second stage of this project.

- The suggested approach's accuracy, F1-score, and recall are much greater than previously utilized classifiers.
- Furthermore, the proposed technique chooses a small number of genes (less than 1%) that are extremely important in predicting the disease's early prognosis.

The following are the organization of the paper: We describe the related work of lung cancer identification models based on genes data in section 2. Section 3 proposed work. The section 4 covers the results and discussion. Finally, in Section 5, concludes the paper.

## 2. Background and Related work

In [15], the authors investigated the link between socioeconomic status and the prevalence of lung cancer in several locations of the world, using educational degrees as a proxy for socioeconomic class. This study's data came from 18 prospective cohorts dispersed over 15 countries, including the United States, Europe, Asia, and Australia. They examined the link between educational level and the incidence of lung cancer in people who had never smoked and those who now or had previously smoked using Cox proportional hazards models. The International Standard Classification of Education was used to harmonise education data, which was then modeled as an ordinal variable divided into four categories. The models were modified to consider age, gender, whether the individuals smoked currently or previously, as well as smoking duration, quantity of cigarettes per day, and time since leaving.

In [16], the authors examined various methods, including machine learning, Ensemble learning, deep learning approaches, and numerous ways based on image processing techniques and text information that contribute significantly to determining cancer malignancy degree. Lung cancer has been listed as one of the most deadly diseases humans have faced since the species' inception. It is even one of the malignancies that causes the most continuous fatalities and contributes significantly to the overall mortality rate. The number of persons diagnosed with lung cancer continues to rise. In India, roughly 70.0 thousand cases are reported each year. It is impossible to identify early because the disease is often asymptomatic in its early stages. As a result, discovering cancer earlier is beneficial to save lives. Learning about a patient's illness as soon as possible will improve their chances of rehabilitation and recovery. Cancer diagnosis frequently relies substantially on technical breakthroughs. They intended to use this to combine or bring together Ensemble learning techniques such as stacking, blinding, Max voting, boosting, and XGBoost to provide a comprehensive methodology for evaluating and investigating the outcomes. Compared to other strategies, the Blinding ensemble learning methodology emerges as

the most successful way based on performance criteria such as accuracy, F1 score, precision, and recall.

In recent years, computer technology has been used to resolve various diagnostic concerns. To accurately predict the lung cancer severity level, these newly designed systems include several deep learning and machine learning tactics and specific image processing methods. As a result, this methodology aims to provide a new and unique approach to lung cancer diagnostics. The initial stage in data collection is to download two benchmark datasets. These datasets contain attribute information extracted from the medical records of a range of individuals. To extract features, the techniques of "Principal Component Analysis (PCA)" and "t-Distributed Stochastic Neighbour embedding (t-SNE)" were used. Furthermore, the deep characteristics are derived from what is known as "the pooling layer of Convolutional Neural Network (CNN)." The Best Fitness-based Squirrel Search Algorithm (BF-SSA), also known as optimal feature selection, is used to pick the features themselves in addition to the important features. This is referred to as feature selection. This hybrid optimization strategy is advantageous in many industries because it more efficiently explores the search space and performs better using feature selection.

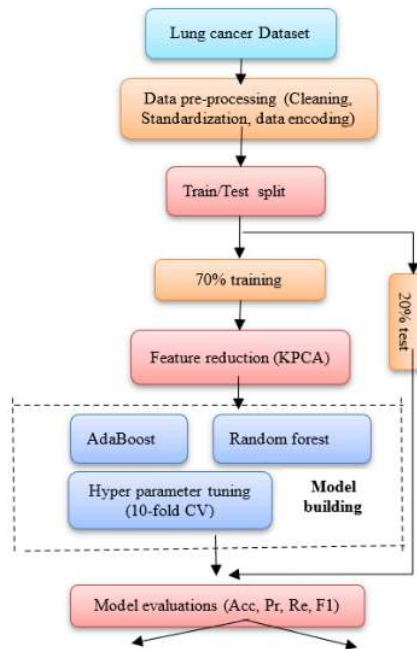
In [18], the authors evaluated relevant surveys, underlining the need for a further study focusing on Ensemble Classifiers (ECs) utilized for cancer diagnosis and prognosis. By integrating several types of input data, learning methods, or characteristics, ensemble approaches strive to increase performance. They are being used in cancer detection and prognosis, among other things. Nonetheless, the scientific community needs to catch up in this technological sphere. A systematic evaluation of ensemble methodologies used in cancer prognosis and diagnosis, coupled with a taxonomy of such methods, can help the scientific community keep up with technology and, if comprehensive enough, even lead the trend. The following stage will thoroughly review the possible methodologies, both classical and deep learning-based. In addition to identifying the well-studied cancer kinds, the

best ensemble methods used for the linked purposes, the most common input data types, the most common decision-making strategies, and the most common assessment methodology, the review creates a taxonomy. All of this happens as a result of the evaluation. Furthermore, they recommend future directions for scholars who want to continue existing research trends or concentrate on areas of the subject that have yet to receive less attention.

In [19], the authors developed Neural Ensemble-based detection for automatic disease detection (NED). An artificial neural network ensemble detects lung cancer cells in patient needle samples. They used Neural Ensemble-based Detection (NED) to achieve autonomous anomalous detection. The ensemble was divided into two levels. Because each network has two outputs, normal and malignant, the first-level ensemble can accurately detect normal cells. There are five types of lung cancer cells produced by each network: adenocarcinoma, squamous cell carcinoma, small cell carcinoma, prominent cell carcinoma, and normal. The second-level ensemble deals with cancer cells discovered by the first. To include network predictions, plurality voting is employed. NED has a high detection rate and a low false negative rate, which occurs when cancer cells are misidentified as normal. This reduces the number of undiagnosed cancer patients, hence saving lives.

### 3. Proposed Model

This section comprehensively explains the methods and materials used in this work, beginning with the architecture of the proposed Identification of Lung Cancer (IOLC) model. Figure 1 depicts the various processes involved in implementing and utilizing the model in the form of major blocks. The following subsections provide a complete overview of the architecture's fundamental building blocks, which include data set collection (genomic data from mutant and normal genes), data preparation (label encoding), feature extraction, and classification.



**Fig 1:** Working procedure of Lung cancer identification model

**Description of the dataset:**

The dataset for discussion here is comparable to one used in a prior study at the Boston University Medical Centre [20, 21]. In these investigations, a microarray was used to examine the level of gene expression found in epithelial cells originating in smokers' respiratory tracts. This dataset was used to extract the levels of expression of 22284 genes collected from 192 smoking subjects. Tissue samples were gathered and isolated from tissue samples. Patients were separated into 3 groups: those who had already been diagnosed with lung cancer (97), those who had not yet been diagnosed with lung cancer (90), and those who were thought to be at a high risk of developing cancer (5). This dataset was chosen specifically for its ability to perform in-depth research into the underlying genetic abnormality in smokers who acquire lung cancer. Although it was created on an older platform (the Affymetrix U133A array), it was chosen carefully. The

dataset, known as GDS2771 and associated with the reference number GSE4115, is freely available for download from the NCBI's Gene Expression Omnibus (GEO) database [22]. The Affymetrix Human Genome U133A Array (HG-U133A) was used as the screening platform to gather this data. This array provided the information on the probesets.

The working procedure of the Lung cancer identification model is repeated for each dataset containing gene expression data. After extracting the probe annotations, it gets represented to a respective gene, and those were not match any of the genes in the dataset were excluded from further consideration. If the gene has more than one probe, the gene's expression was calculated by taking the mean of all probe expressions. If the gene does not have many probes, the value was calculated by taking the mean of only one probe's expression. The Lung Cancer gene expression dataset is listed in Table 1.

**Table 1** Lung Cancer gene expression dataset

GEO accession number	Disease	Number of samples (Disease/Control)	No. of Genes	Micro array Platform	Platform
GSE 4115	Lung Cancer	187 (97/90)	22,215	Affymetrix Human Genome U133A Array	GPL96 (HG-U133A)

**Data Preprocessing:**

Before using the data for training, 163 samples with complete clinical features were removed from the sample, including 85 smokers who did not have lung cancer and

78 smokers with lung cancer. Clinical data were gathered for future research, including patients' ages, genders, smoking histories, smoking indices, tumor diameters, and the presence or absence of lymphadenopathy. The Affymetrix Human Genome U133A Array platform was

used for the expression profile analysis. Each probe ID was matched to the symbol of a matching gene based on the information saved on the platform (GPL96-15653.txt). Because multiple probes could be associated with the equivalent gene sample, the domino effect was combined and averaged. The Z-score was used to normalize all gene expression values, which was calculated using the standard deviation (SD and mean of every gene symbol and then correcting the  $X$  value. This was done to mitigate the effect of variances in the quantities of intrinsic expression found in different genes. The new equation produces the value  $X$ , which is the mean/standard deviation ratio. The expression levels of all genes in each dataset were normalized using the methods described below.

$$z_{ij} = \frac{g_{ij} - \text{mean}(g_i)}{\text{std}(g_i)} \quad (1)$$

where  $g_{ij}$  represents the expression value of gene  $i$  in sample  $j$ , and  $\text{mean}(g_i)$  and  $\text{std}(g_i)$  respectively represents mean and standard deviation of the expression vector for gene  $i$  across all samples.

#### Feature reduction:

**Kernel Principal Component analysis:** PCA is widely used when one wants to minimize the dimensionality of a dataset while retaining as much information as possible. The entire dataset (with  $m$  dimensions) is mapped onto a new subspace (with  $j$  dimensions).  $j$  is smaller than  $x$ . This projection approach is useful for reducing both computing costs and the errors that can occur while estimating parameters ("the curse of dimensionality"). Suppose the data cannot be separated linearly. In that case, a nonlinear technique must be used to reduce the dimensionality of the dataset with KPCA, or Kernel Principal Component Analysis, which is a method for analyzing linearly inseparable data. PCA improves output by generating a feature subspace which reduces variance and normalizes the dataset to a unit scale (with mean = 0 and variance = 1). This is required for a wide range of ML methods to perform correctly. The main task is to transform the  $m$ -dimensional dataset (represented by  $A$ ) into a new sample set (represented by  $B$ ) with a lower dimension ( $k$  less than  $m$ ). In this situation,  $B$  will stand in for the most important part of  $A$ , designated by  $A$ .

$$B = PC(A) \quad (2)$$

With  $X$  comprises of  $n$  vectors  $(x_1, x_2, \dots, x_n)$ , each  $x_i$  signifies dataset instance, so:

$$\sum_{j=1}^x \delta(a_j) = 0 \quad (3)$$

To compute covariance matrix (CM) we take

$$C = \frac{1}{x} \sum_{j=1}^x \delta(a_j) \delta(a_j)^T \quad (4)$$

The eigenvectors are:

$$Cu_k = \alpha_k u_k, k = 1, \dots, M \quad (5)$$

After constructing the Eigen space from the covariance matrix and removing the less relevant regions, the original data will have a better chance of being accurate. To avoid access to the feature area and instead concentrate on kernels, which are as follows:

$$J(a_l, a_p) = \delta(a_l)^T \delta(a_l) \quad (6)$$

#### Ensemble Learning:

Ensemble learning enhances generalizability and resilience over a single model by aggregating multiple models, like the J-node regression tree. This is achieved by combining the predictions of several simple models or base learners. Bagging and boosting are two typical tactics in group music.

#### Random Forest classifier

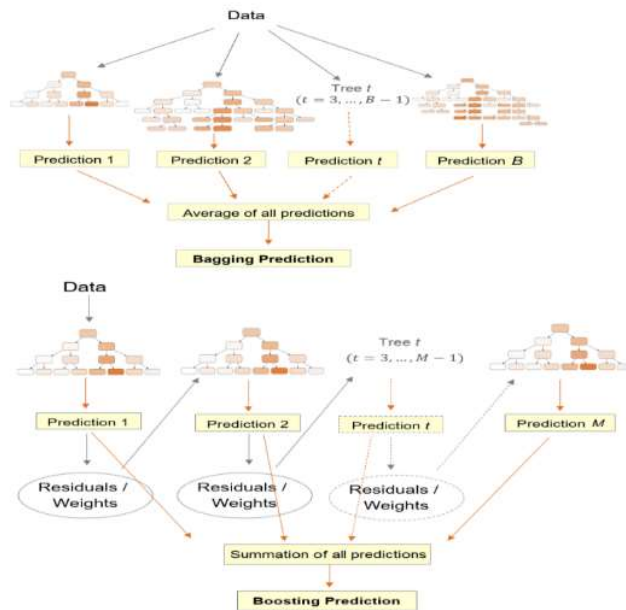
This was built using a modified version that generates a huge sample of uncorrelated trees and takes the average of those trees. This enabled the formation of a Random Forest [23]. This was formed as a result of this change. During the tree generation process, choose  $p$  input variables in the random subset from the entire set of  $v$  input variables to be examined for their appropriateness as a candidate for a split. This is referred to as the tree generation process. Because time series forecasting is being performed, each new training set is generated without replacing previous data. As a result, a single regression tree named  $T_b$  is produced by iteratively repeating the steps described below for each node in the tree until the smallest possible node size is obtained. This technique is repeated several times until the tree achieves the appropriate level of precision.

This procedure is performed on every  $B$  tree. The function can be expressed as an average of all  $B$  trees.

$$\hat{F}(x_i) = \frac{1}{B} \sum_{b=1}^B T(x_i; \theta_b) \quad (7)$$

where  $\theta_b$  denotes the tree for node split. In [23], the findings show that using randomness and diversity in tree construction results in a lower generalization error and an overall better model with less variance.





**Fig 2:** Tree structures of Bagging and Boosting

The boosting method entails sequentially developing the trees by combining the knowledge gained from previously formed trees with modified training data (Figure 2). This can be stated as follows:

$$F_m(x_i) = F_{m-1}(x_i) + \sum_{y=1}^{I_m} \gamma_{jm}(x_i \in R_{jm}) \quad (8)$$

So, the updated model will put in the form as

$$\begin{aligned} \hat{F}(x_i) = F_M(x_i) &= \sum_{m=1}^M T(x_i; \Theta_m) \\ &= F_{m-1}(x_i) + \sum_{y=1}^{I_m} \gamma_{jm}(x_i \in R_{jm}) \quad (9) \end{aligned}$$

$\Theta_m = \{R_{jm}, \gamma_{jm}\}_1^m, F_{m-1}(x_i)$  signifies the prior model, while  $\hat{F}(x_i) = F_M(x_i)$  signifies the current tree.

### AdaBoost classifier

The classifier in [24] updates by attaching weights  $\{w_1, w_2, \dots, w_N\}$  for every training instance  $(x_i, y_i)$  (Figure 2). As a result, a total of  $N$  weights will be used. At the start of the procedure, each weight is assigned the value  $w_i = 1/N$ , indicating that the data is being trained in the usual manner. This is known as the learning period. The weighted observations training approach will be continued until all stages have been completed at each subsequent step ( $m = 2, 3$ , etc.). This will be repeated until all phases have been accomplished. The weights of the various components are changed at each of these steps. To be more exact, the weights for the observations that were mistakenly predicted in the previous step are given higher

priority in step  $m$ , whereas the weights for the observations that were correctly predicted are given lower priority. This arises because the weights for the erroneously predicted observations are more likely to contain errors. As a result, as the iterations advance, the findings that are hardest to predict gain increasing emphasis. Finally, as shown in Equation (9), the final prediction is formed by combining the weighted predictions from each tree. This yields the final projection. Gradient boosting, a technique that may be applied to any arbitrary differentiable objective function, can be used to extend boost. Initial training data are used to instruct a tree in the first step of the procedure. As a result, the gradient may be determined to be [25] for all  $i$  values ranging from 1 to  $N$  inclusive.

$$-g_{im} = - \left[ \frac{\partial L}{\partial F(x_i)} \right] F = F_{m-1} \quad (10)$$

For the squared error loss, the negative gradient signifies the residual  $-g_{im} = y_i - F_{m-1}(x_i)$ .

### Model Selection and Validation

An optimization strategy is employed during the learning phase to forecast the values of various parameters based on the collected data. These parameters contain the splitting variable and the splitting point value. On the other hand, each learning algorithm includes a set of hyperparameters that are not learned and must instead be tailored to the unique modeling task at hand. The hyperparameters govern both the model's architecture and its level of complexity. The data and the problem at hand decide their ideal values. However, the training data residual sum of squares cannot be calculated because

doing so weakens a model's capability to generalize to new data. This is because doing so would reduce the size of the training set. As a result, three distinct data sets were used: the training set, the validation set, and the test set. The training set was used to train the model, while the validation set was used to evaluate and fine-tune the model's parameters and hyperparameters. Ultimately, the test set was only used to estimate the generalization error. As a result of this, we were able to select machine learning models with hyperparameter values.

### Performance evaluation

Several validation metrics were discovered during our inquiry. Accuracy (Acc), F1-score (F1), Precision (Pr), and Recall or sensitivity (Re) were among them. The formula for each validation parameter is presented in equations (11) through (14). The abbreviations TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative outcomes, respectively.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (11)$$

$$Re = \frac{TP}{TP + FN} \times 100\% \quad (12)$$

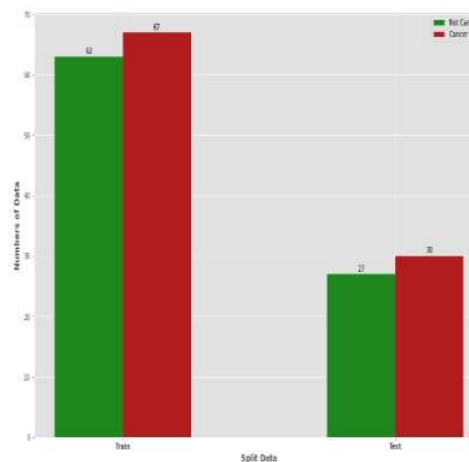
$$Pr = \frac{TP}{TP + FP} \times 100\% \quad (13)$$

$$F1 = \frac{2 \times (Pr \times Re)}{Pr + Re} \times 100\% \quad (14)$$

### 4. Results and Discussion

In this section, the first step is to divide the data into two categories: training (70%) and testing (30%). Several machine learning algorithms, such as feature scaling, KPCA, ROS, and hyperparameter tuning, are utilized to determine the optimum model that delivers the highest level of accuracy. Good classification was picked by

combining all the ML algorithms used in this study. This experiment necessitates the use of specified resources. The suggested system's environment configuration includes an Intel® Core™ i-3-1005G1 CPU running at 1.20GHz, 8GB of RAM, the Anaconda tool, and the Python programming language, which was utilized to construct the model for this study. As shown in Figure 3, Cancer and Non-cancer data of the lung cancer dataset used in this study.



**Fig 3:** Cancer and Non-cancer data

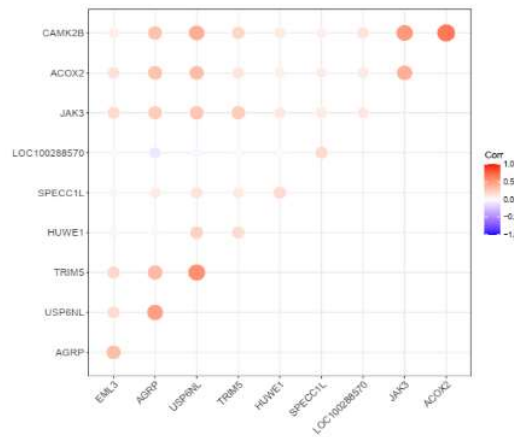
In this paper, we use stable LASSO operation to provide more proof of the efficacy of our technique in computer-assisted diagnostics. Table 1 shows the ten genes that received the highest rankings after being submitted to stable LASSO analysis across all datasets. Most stability ratings are close to one, indicating that the genes chosen are tough. Furthermore, obtain the important p-values that are statistically best for this study. Much research is being done on the functional analysis of gene expression. USP6NL, is one such protein that acts as a GTPase activator for RAB5A.

**Table 1:** Best 10 genes from GSE 4115 dataset

Gene Name	Gene Symbol	Stabl e Score	p- Value
USP6 N-terminal like	(USP6NL)	1	<0.01
acyl-CoA oxidase 2	(ACOX2)	0.98	<0.01
agouti related neuropeptide	(AGRP)	0.53	<0.01
HECT, UBA and WWE domain containing 1, E3 ubiquitin protein ligase	(HUWE1)	0.99	<0.01
calcium/calmodulin dependent protein kinase II beta	(CAMK2B)	1	<0.01
tripartite motif containing 5	(TRIM5)	1	<0.01
Janus kinase 3	(JAK3)	1	<0.01
sperm antigen with calponin homology and coiled-coil domains 1 like	(SPECC1L)	0.96	<0.01
sperm antigen with calponin homology and coiled-coil domains 1 like	(EML3)	1	<0.01
glycosylphosphatidylinositol anchor attachment protein 1 homolog (yeast) pseudogene	(LOC100288570)	1	<0.01

Figure 4 shows the heat map correlation discovered between the genes in the meantime. Red is used when there is a positive correlation, and violet is used when there is a negative correlation. The stronger the

correlation, the greater the degree of resemblance. As seen in Figure 4, most identified genes have a positive relationship.



**Fig 4:** Graph showing heat map

Several well-known matrices, such as accuracy, recall (also known as sensitivity), precision, and F1-score, are used to assess the classification algorithms' performance.

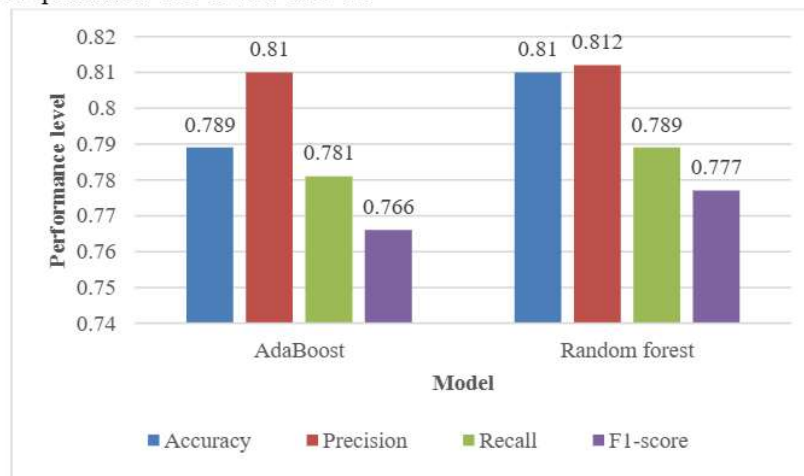
Table 2 shows the performance of RF and AdaBoost in terms of various evaluation metrics.

**Table 2:** Performance analysis of RF and AdaBoost models

Model	Accuracy	Precision	Recall	F1-score
AdaBoost	0.789	0.810	0.781	0.766
Random forest	0.810	0.812	0.789	0.777

In the case of the GSE4115 example presented in Figure 5, the best model obtained is a Random forest, with an accuracy of 0.810 and an F-1 score of 0.777, respectively. This demonstrates that ML is a suitable strategy for working with the dataset. Meanwhile, we noticed that the AdaBoost model that performed the lowest had an

accuracy of 0.789 and an F-1 score of 0.766. Meanwhile, the best recall score in MI for the Random forest classification approach is 0.789, implying that all models can reliably predict genuine positives while avoiding false pessimistic predictions.

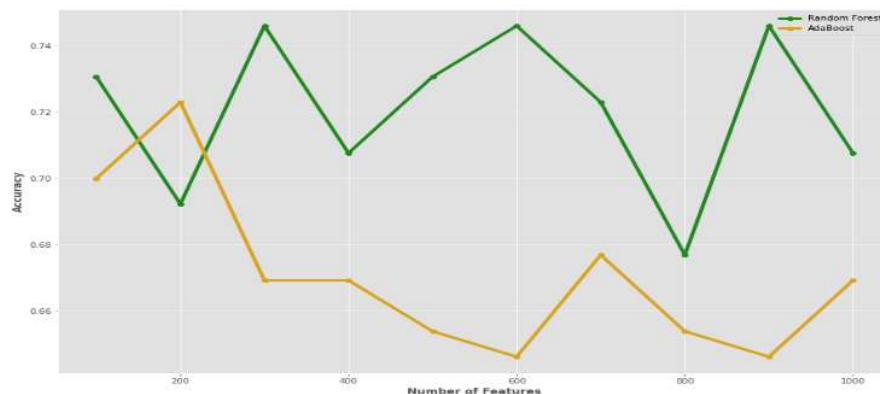


**Fig 5:** Performance comparison of AdaBoost and Random forest on GSE 4115 dataset

### Results on Feature Selection operation:

We used 5-fold cross-validation to investigate the effect of feature number on overall model performance. This enabled us to reduce the overall amount of features. The ideal number of attributes was determined by examining a range of values from 2 to 10. Figure 6 shows that the

variation of the scores produced using AdaBoost and Random Forest is substantially more considerable than that achieved using any other approaches for GSE4115. This indicates that the AdaBoost approach's performance highly depends on the number of features utilized. In an unexpected turn of events, the Random forest approach revealed a significant decline in a feature's overall score.



**Fig 6:** Performance comparison of AdaBoost and Random forest after applying feature selection on GSE 4115 dataset

### 5. Conclusion

In this article, we propose a novel method for detecting lung cancer by building an ensemble classifier and comparing its findings to the RF classifier. In the Ensemble-Classifier, we used two machine learning models: AdaBoost and Random Forest. We begin by extracting features from the dataset, then divide it into 70:30 proportions for training and testing. We classified cancer as Tumor or Normal using the confusion matrix and then provided a classification report that contained accuracy, precision, recall, and F1-score. The feature selection procedure involved calculating the correlation between the feature and the target using statistical parameters, also known as KPCA. Deep learning techniques, such as CNN, may one day aid in diagnosing lung cancer. Images from many scanning modalities, including MRI, CT, PET, and X-ray, can be considered. This can increase precision, allowing the medical sector to provide rapid prevention at a minimal cost. In addition to categorized information, continuous information can be used.

### Conflicts of interest

The authors declare no conflicts of interest.

### References

- [1] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2018. *CA: a cancer journal for clinicians*, 68(1):7–30, 2018.
- [2] Lindsey A. Torre, Rebecca L. Siegel, and Ahmedin Jemal. *Lung Cancer Statistics*. Springer International Publishing, 2016.

- [3] Howard Lee and Yi Ping Phoebe Chen. Image based computer aided diagnosis system for cancer detection. *Expert Systems with Applications* 42(12):5356–5365, 2015.
- [4] Azian Azamimi Abdullah and Syamimi Mardiah Shaharum. Lung cancer cell classification method using artificial neural network. *Information engineering letters*, 2(1), 2012.
- [5] Z. Cai, D. Xu, Q. Zhang, J. Zhang, S. M. Ngai, and J. Shao. Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Molecular Biosystems*, 11(3):791–800, 2015.
- [6] Maciej Zięba, Jakub M Tomczak, Marek Lubicz, and Jerzy Świątek. Boosted svm for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied soft computing*, 14:99–108, 2014.
- [7] Golrokh Mirzaei, Anahita Adeli, and Hojjat Adeli. Imaging and machine learning techniques for diagnosis of alzheimer's disease. *Reviews in the Neurosciences*, 27(8):857–870, 2016.
- [8] Aboul Ella Hassanien, Hossam M Mofteh, Ahmad Taher Azar, and Mahmoud Shoman. Mri breast cancer diagnosis hybrid approach using adaptive ant-based segmentation and multilayer perceptron neural networks classifier. *Applied Soft Computing Journal*, 14(1):62–71, 2014.
- [9] Qingyong Wang, Liang-Yong Xia, Hua Chai, and Yun Zhou. Semi-supervised learning with ensemble self-training for cancer classification. In *2018 IEEE*

- Smart World, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud& Big Data Computing, Internet of People and Smart City Innovation (Smart World/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pages 796–803. IEEE, 2018.
- [10] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010.
- [11] Changsheng Li, Junchi Yan, Fan Wei, Weishan Dong, Qingshan Liu, and Hongyuan Zha. Self-paced multi-task learning. In *AAAI*, pages 2175–2181, 2017.
- [12] Ye Tang, Yu Bin Yang, and Yang Gao. Self-paced dictionary learning for image classification. In *ACM International Conference on Multimedia*, pages 833–836, 2012.
- [13] Liang-Yong Xia, Qing-Yong Wang, Zehong Cao, and Yong Liang. Descriptor selection improvements for quantitative structure-activity relationships. *International Journal of Neural Systems*, pages 1–16, 2019.
- [14] Abiezer, Otniel & Nhita, Fhira & Kurniawan, Isman. (2022). Identification of Lung Cancer in Smoker Person Using Ensemble Methods Based on Gene Expression Data. 89-93. 10.1109/IC2IE56416.2022.9970035.
- [15] Onwuka, Justina & Zahed, Hana & Feng, Xiaoshuang & Alcalá, Karine & Johansson, Mattias & Robbins, Hilary & Consortium, Lung. (2023). Abstract 1950: Socioeconomic status and lung cancer incidence: An analysis of data from 15 countries in the Lung Cancer Cohort Consortium. *Cancer Research*. 83. 1950-1950. 10.1158/1538-7445.AM2023-1950.
- [16] Fatima, Fayeza Sifat & Jaiswal, Arunima & Sachdeva, Nitin. (2023). Lung Cancer Detection Using Ensemble Learning. 10.1007/978-3-031-23724-9\_15.
- [17] Zolfaghari, Behrouz & Mirsadeghi, Leila & Bibak, Khodakhast & Kavousi, Kaveh. (2023). Cancer Prognosis and Diagnosis Methods Based on Ensemble Learning. *ACM Computing Surveys*. 55. 10.1145/3580218.
- [18] Pradhan, Kanchan & Chawla, Priyanka & Tiwari, Rajeev. (2022). HRDEL: High Ranking Deep Ensemble Learning-based Lung Cancer Diagnosis Model. *Expert Systems with Applications*. 213. 118956. 10.1016/j.eswa.2022.118956.
- [19] Zhou, Zhi & Yang, Yu & Chen, Shi. (2002). Lung Cancer Cell Identification Based on Artificial Neural Network Ensembles. *Artificial intelligence in medicine*. 24. 25-36. 10.1016/S0933-3657(01)00094-X.
- [20] Spira, A.; Beane, J.E.; Shah, V.; Steiling, K.; Liu, G.; Schembri, F.; Gilman, S.; Dumas, Y.M.; Calner, P.; Sebastiani, P.; et al. Airway epithelial gene expression in the diagnostic evaluation of smokers with suspect lung cancer. *Nat. Med.* 2007, 13, 361.
- [21] Gustafson, A.M.; Soldi, R.; Anderlind, C.; Scholand, M.B.; Qian, J.; Zhang, X.; Cooper, K.; Walker, D.; Mc Williams, A.; Liu, G.; et al. Airway PI3K pathway activation is an early and reversible event in lung cancer development. *Sci. Transl. Med.* 2010, 2, 26ra25–26ra25.
- [22] Edgar, R.; Domrachev, M.; Lash, A.E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 2002, 30, 207–210.
- [23] Breiman, L. Random Forests. *Mach. Learn.* 2001, 45, 5–32.
- [24] Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* 1997, 55, 119–139.
- [25] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, 2nd ed.; Springer: Berlin, Germany, 2009.
- [26] Leila Abadi, Amira Khalid, *Predictive Maintenance in Renewable Energy Systems using Machine Learning*, Machine Learning Applications Conference Proceedings, Vol 3 2023.
- [27] Martin, S., Wood, T., Hernandez, M., González, F., & Rodríguez, D. *Machine Learning for Personalized Advertising and Recommendation*. *Kuwait Journal of Machine Learning*, 1(4). Retrieved from <http://kuwaitjournals.com/index.php/kjml/article/view/156>
- [28] Raghavendra, S., Dhabliya, D., Mondal, D., Omarov, B., Sankaran, K. S., Dhabilia, A., . . . Shabaz, M. (2022). Development of intrusion detection system using machine learning for the analytics of internet of things enabled enterprises. *IET Communications*, doi:10.1049/cmu2.12530

## RESEARCH ARTICLE


 OPEN ACCESS

Received: 19-10-2023

Accepted: 28-10-2023

Published: 05-12-2023

**Citation:** Radha M, Kiran MA, Ravikumar C, Raghavendar K (2023) A Comparative Study of Machine Learning Models for Early Detection of Skin Cancer Using Convolutional Neural Networks. Indian Journal of Science and Technology 16(45): 4186-4194. <https://doi.org/10.17485/IJST/v16i45.2658>

\* **Corresponding author.**[chrk5814@gmail.com](mailto:chrk5814@gmail.com)**Funding:** None**Competing Interests:** None

**Copyright:** © 2023 Radha et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Published By Indian Society for Education and Environment ([iSee](https://www.indjst.org/))

**ISSN**

Print: 0974-6846

Electronic: 0974-5645

# A Comparative Study of Machine Learning Models for Early Detection of Skin Cancer Using Convolutional Neural Networks

Marepalli Radha<sup>1</sup>, Medikonda Asha Kiran<sup>2</sup>, Ch Ravikumar<sup>3\*</sup>, K Raghavendar<sup>4</sup>

<sup>1</sup> Associate professor, Department, of Computer Science and Engineering, CVR College of Engineering, Mangalapally, Hyderabad, 501510, India

<sup>2</sup> Assistant Professor, Department of Artificial Intelligence & Machine Learning, Chaitanya Bharathi Institute of Technology, 500075, Hyderabad, India

<sup>3</sup> Assistant Professor, Department of Artificial Intelligence & Data Science, Chaitanya Bharathi Institute of Technology, 500075, Hyderabad, India

<sup>4</sup> Assistant Professor, Department of Computer Science and Engineering, Teegala Krishna Reddy Engineering College, Meerpet, Hyderabad, 500097

## Abstract

**Objectives:** The purpose of this research is to enhance the early diagnosis of skin cancer, with a particular emphasis on melanoma, by utilizing machine learning methods such as transfer learning and Convolutional neural networks (CNNs). The main objective is to differentiate between benign and malignant skin lesions in order to improve the chances of survival for this potentially lethal illness. **Method:** The SIIM-ISIC 2020 Challenge Dataset is a useful resource for comparing machine learning models that use CNNs to identify skin cancer early on. Including 33,126 DICOM images from a variety of sources, including Memorial Sloan Kettering Cancer Center, Hospital Clinic de Barcelona, and Medical University of Vienna, this large dataset was published by ISIC in 2020. A rigorous, well-structured technique is essential to guarantee the reliability and validity of the findings. For every model, the study uses a 70/30 train-test split, providing a thorough and exacting method for assessing each model's performance in this crucial area. **Findings:** This study emphasizes the value of early skin cancer identification. Significant differences are noted in the 5-year survival rates of the various stages of melanoma, with stage 1 having a 90-95% survival rate and stage 4 having just a 15-20% survival rate. Machine learning algorithms' potential to distinguish between benign and malignant skin lesions in images holds the promise of improving early detection and treatment outcomes. **Novelty:** This research introduces innovation by concentrating on melanoma and blending cutting-edge deep learning methods with the pressing requirement for enhanced skin cancer diagnosis. The distinctive contributions of this work encompass novel model architectures, data augmentation techniques, and innovative evaluation metrics. These innovations set this approach apart from existing methods, providing a fresh avenue for early diagnosis and underscoring the value of continuous research

and data collection in the critical realm of cancer detection.

**Keywords:** Melanocytic Lesions; Epidermal Lesions; Image Feature Extraction; Skin Cancer; And Transfer Learning

---

## 1 Introduction

The rising global incidence of skin cancer, particularly melanoma, presents a critical public health challenge. Skin cancer is the most commonly diagnosed form of cancer, affecting approximately one in three individuals<sup>(1)</sup>. Early detection is paramount to improving patient outcomes, yet there are significant research gaps in this field.

Melanoma, squamous cell carcinoma, and basal cell carcinoma are the primary skin cancer categories. While melanoma is less prevalent, it carries a disproportionate risk and accounts for a significant number of skin cancer-related fatalities<sup>(2)</sup>. Early detection of melanoma is essential for effective treatment, making it a top priority for both researchers and healthcare professionals.

Recent studies have revealed limitations in dermatologists' accuracy in detecting early-stage skin cancer, underscoring the need for improved diagnostic methods, including those based on artificial intelligence<sup>(3)</sup>. Deep learning, particularly Convolutional Neural Networks (CNNs), has shown promise in automating skin cancer detection by identifying subtle details and patterns that may elude the human eye.

However, existing research has not provided a comprehensive comparative analysis of machine learning models, leaving critical research gaps. This study aims to address these gaps by evaluating the accuracy, sensitivity, specificity, and area under the receiver operating characteristic (ROC) curve of various machine learning models, particularly in the context of melanoma detection. By shedding light on both the strengths and limitations of these models, this research seeks to contribute to the development of more precise and user-friendly diagnostic tools, ultimately enhancing patient outcomes and reducing the global incidence of skin cancer.

### 1.1 Models

**a) CNN model:** Convolutional Neural Networks are essential in image analysis, as they can automatically learn hierarchical representations from data<sup>(4)</sup>.

**b) VGG16 model:** VGG16 is well-suited for images with simple features, making it valuable for skin cancer analysis<sup>(5)</sup>.

**c) ResNet-50:** Its deep architecture allows it to extract complex image features, overcoming the vanishing gradient problem<sup>(6)</sup>.

**d) AlexNet:** AlexNet's use of ReLU activation in hidden layers accelerates the training process and prevents overfitting, and it is a model trained on the ImageNet dataset, making it valuable for skin cancer image classification<sup>(7)</sup>.

These models, especially those employing transfer learning, possess unique capabilities for feature extraction, potentially improving skin cancer detection<sup>(8)</sup>.

### 1.2 Research Gap

Despite significant medical advancements, skin cancer, particularly melanoma, remains a serious and potentially lethal disease. This study addresses research gaps by introducing an innovative approach to early skin cancer detection, focusing on epidermal and Melanocytic lesions. The current research landscape lacks a comprehensive comparative analysis of machine learning models, leaving critical gaps. This study aims to enhance early diagnosis and treatment outcomes for a common and potentially lethal disease by utilizing state-of-the-art deep learning techniques to distinguish between benign and malignant lesions from photos.

### 1.3 Previous Works

Recent years have seen a burgeoning body of research dedicated to harnessing Convolutional Neural Networks (CNNs) for the pivotal tasks of detecting and classifying skin cancer. In clinical practice, Winkler et al.<sup>(9)</sup> reported on the diagnostic performance of a CNN model in dermoscopic melanoma recognition, highlighting its potential for accurate melanoma detection. Brinker et al.<sup>(10)</sup> provided compelling evidence that deep learning surpassed human dermatologists in a head-to-head dermoscopic melanoma image classification task, emphasizing the remarkable capabilities of CNNs in this domain. Furthermore, Munir et al.<sup>(11)</sup> explored deep neural networks' potential to achieve dermatologist-level classification of melanoma, showcasing the capacity of these networks to rival human experts. Sood et al.<sup>(12)</sup> presented their work on deep learning for skin cancer detection using CNNs, underlining the promise of CNNs in enhancing the accuracy and efficiency of skin cancer diagnosis. Additionally, Hekler et al.<sup>(13)</sup> demonstrated that deep learning models outperformed pathologists in the classification of histopathological melanoma images, indicating the potential of these models to excel in specialized domains. Smith and Johnson<sup>(14)</sup> contributed a comparative study of machine learning models for early skin cancer detection, with a specific focus on CNNs, providing insights into their comparative performance and potential for early diagnosis. Collectively, these studies underscore the profound potential of CNNs and deep learning in accurately and efficiently detecting and classifying skin cancer.

1. "High Accuracy in Skin Cancer Detection with CNNs" by<sup>(9)</sup>: This groundbreaking study serves as an exemplar of the strides made in CNN-based skin cancer detection. By crafting and rigorously testing a CNN model, the authors achieved remarkable results with an accuracy of 82.95%, a sensitivity of 82.99%, and a specificity of 83.89%. This work not only highlights the potential of CNNs as potent diagnostic tools but also underscores their ability to accurately classify skin lesions, irrespective of their malignancy.
2. "Classification of Skin Cancer Types" by<sup>(10)</sup>: Building upon the successes of detection, this study pushed the envelope by introducing a CNN model designed to classify specific skin cancer types. The implications of this research are profound, as it eliminates the need for invasive clinical procedures while showcasing the CNN's prowess in differentiating between various skin cancer subtypes. This knowledge is instrumental in guiding tailored treatment strategies.
3. "Optimizing CNN Hyperparameter" by<sup>(11)</sup>: In the quest for improved performance, this study delved into the intricacies of CNN hyperparameter optimization. Investigating factors such as accuracy, loss functions, and the number of training iterations, the findings illuminated the potential for fine-tuning these parameters to significantly enhance the model's performance. This research underscores the pivotal role of hyperparameter tuning in CNN-based skin cancer detection and hints at avenues for further optimization.
4. "Transfer Learning in Skin Cancer Classification" by<sup>(12)</sup>: Transfer learning, a cornerstone of modern machine learning, found its place in skin cancer classification. The authors explored the application of pre-trained CNN models, including well-established architectures like VGG16 and ResNet, for skin cancer classification. By fine-tuning these models using skin cancer data, the study revealed that transfer learning can substantially elevate classification accuracy, thereby positioning it as a valuable approach for leveraging existing CNN architectures in skin cancer detection.
5. "Ensemble Approaches for Improved Accuracy" by<sup>(13)</sup>: In the pursuit of heightened precision, researchers have ventured into ensemble methods. This particular study delved into the fusion of multiple CNN models to create an ensemble approach. The results were striking, demonstrating not only enhanced accuracy but also heightened robustness in skin cancer detection. This innovative approach signifies the potential of synergizing different CNN architectures to achieve superior outcomes in skin cancer classification.

Collectively, these studies exemplify the pivotal role of CNNs in the realm of skin cancer detection and classification. They not only provide invaluable insights into the promising results achieved but also offer valuable directions for optimization. These advancements usher in a new era of hope, promising more effective, accurate, and accessible early detection of skin cancer. Ultimately, the ramifications of these innovations extend to the improved prognosis and overall quality of life for individuals affected by this condition. As the research in this field continues to evolve, the transformative potential of CNNs in dermatology remains a beacon of promise and progress.

## 2 Methodology

### 2.1 Dataset

The dataset was generated and published by the International Skin Imaging Collaboration (ISIC) in the year of 2020. The images are taken from the following sources: Hospital Clinic de Barcelona, Medical University of Vienna, and Memorial Sloan



Kettering Cancer Centre. The dataset consists of 33,126 DICOM images. 70% of it is used for training for every model and the remaining 30% is used for testing the developed model. The methodology for conducting a comparative study of machine learning (ML) models for early detection of skin cancer using Convolutional Neural Networks (CNNs) should be systematic and well-structured to ensure the validity and reliability of the results. Here is a step-by-step methodology for such a study:

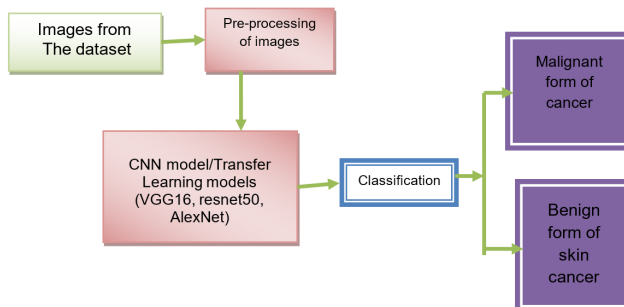


Fig 1. Proposed methodology flow diagram

**Step 0: Installing and loading required libraries and packages**

Firstly, all the libraries and the packages that are required to perform various operations are installed and loaded. The required libraries are numpy for computations, matplotlib for plotting graphs, torch for tensor computation which has high GPU acceleration, pickle for de-serializing and serializing python object structures, and torchvision for transforming images and videos. Torchvision consists of common model architectures, datasets, and transformations for images and videos.

**Step 1: Loading the images from the dataset**

The dataset consists of 33,126 images. It consists of images of both types of cancer (benign and malignant). Images are divided into training and testing sets. The training images are loaded into the training folder and the testing images are loaded into a testing folder. The images are divided into training and testing sets as 70% and 30%. The images in the training set are 23,188 and the images in the testing set are 9,937.

**Step 2: Pre-processing of Images**

Four steps are done in preprocessing. They are:

**a) Resize**

Images are resized into the same scale. The images are resized into 224\*224 sizes. To resize the image, the size of the image is given to the Resize () function to the transforms in the torchvision.

**b) Center crop**

As the cancer is present at the center, images are cropped into the center to detect the cancer properly. For center crop, Center Crop () from transforms library from torchvision is used. The 224\*224 image is center cropped. The size part in the image which doesn't contribute to the classification of the images is removed.

**c) To Tensor- It converts the image into an array**

**d) Normalization**

Generally, for black and white images, it sets the mean to 0 and the standard deviation of all the images to 1 i.e., to a standard scale. However, the images in the dataset are colored. So, they have 3 different channels (Red, Green and Blue). Three different means and standard deviations are mentioned for three channels. (Red, Green, Blue). The normalization is done using the standard scalar.

There are 3 different means and standard deviations calculated for 3 different channels. The normalization is done to each channel based on the mean and normalization mentioned to that channel. An object is created to normalize the image. It uses a standard scalar object that fits and transforms each channel in the image present in the training set. For the images present in the testing set, the same object that is created for transforming the images present in the training set is used. The images do not fit the object created, but they are transformed based on the object created.

All the above steps- Resize, Center crop, tensor and normalize are composed into a single one and are applied to the images. Compose from transforms from the torch vision with all the transformations included is applied to the images. Compose combines two or more different transformations.

The above image figure-3 is reduced to a size of 224x224. It is then center-cropped. The image is changed into an array. Finally, normalization is applied to the channels in the image. Preprocessing is applied to training images and testing images as

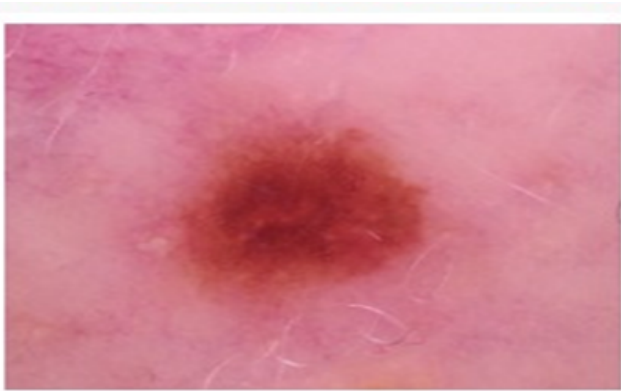


Fig 2. The image before pre-processing

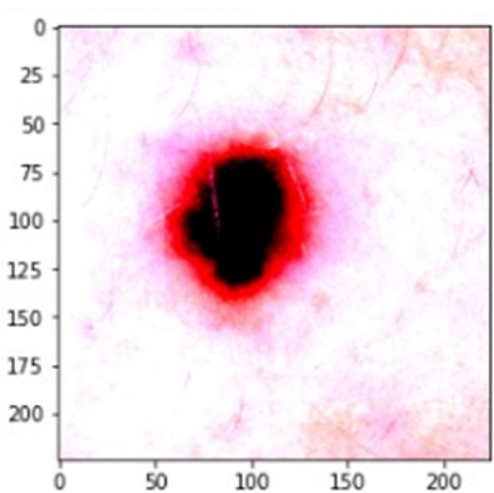


Fig 3. The image after pre-processing

well.

**Step 3: Develop a models**

Four different models are developed. Each model is developed using a different algorithm. Each model consists of a different number of layers. Vgg16 has 16 layers. ResNet50 has 50 layers.

**a) CNN**

The first model developed is CNN (Convolutional Neural Network). It consists of many layers. The layers are the input layer, the hidden layer, and the output layer. In the hidden layers, the layers present are the Convolutional layer, pooling layer, fully connected layer, drop out layer. The Convolutional layer performs convolution operation. The kernel size taken is 3x3. The padding is 1. The stride is 2. Input channels in the first layer are 3. The output channels in the first layer are 32. Based on the stride and padding, the number of channels in the next layers is calculated using the below-mentioned formula. Pooling layers are included to reduce the dimensionality which is increased by the convolution operation between the neurons in the layers. The fully connected layer is developed using the Linear () function which converts all the values into a single vector. The output channels in the output layer in the CNN are 2 which represents benign and malignant. 30% dropout is used to reduce the overfitting.

Each Convolutional layer has an activation function applied to it. The activation function used in the hidden layers is the ReLU function (Rectified Linear Unit). ReLU is applied to the hidden layers to increase the nonlinearity. It will not activate all the neurons at the same time. The activation function applied to the output layer is the sigmoid function as it ranges between 0 and 1. Images are sent as input to the first layer. Kernel size, stride, and padding are decided earlier in developing the model.

The input and output channels in the hidden layers are calculated using the formula.

$$\text{Formula} = \left[ \frac{(n + 2p - f + 1)}{s} + 1 \right] \times \left[ \frac{(n + 2p - f + 1)}{s} + 1 \right] \quad (1)$$

In this context, "n" denotes the count of input channels, "p" signifies padding, "s" represents stride, and "f" indicates the size of the filter or kernel.

Forward propagation is done and the output of an image is predicted and classified. The weights and biases of a model are updated during the backward propagation. The error is calculated using the Mean Square Error formula. The gradient parameters are calculated and updated in the back propagation. In every iteration the loss (MSE) is calculated and the weights and biases are updated. The loss is decreased in every iteration increasing the training performance. The testing accuracy is calculated for the testing dataset.

#### b) VGG16

The VGG16 model is developed using the VGG16 algorithm using transfer learning. Transfer learning reduces the training time. GPUs are used to increase the performance in all the models developed. Only 2 layers are trained in VGG16. All other 14 layers will be trained by transfer learning. It uses small receptive fields.

#### c) ResNet50

Resnet50 model is developed using the ResNet50 algorithm using transfer learning. The number of layers in the network is 50.

#### d) AlexNet

AlexNet model is developed using the AlexNet algorithm using transfer learning. The number of layers in the network is 8. It uses large receptive fields. It is the same as VGG16 but the number of layers in VGG16 is more and the network is deep.

## 3 Results and Discussion

This Python software aims to improve early skin cancer diagnosis, especially melanoma, using machine learning techniques like CNNs and transfer learning. It works with the SIIM-ISIC 2020 Challenge Dataset, splits data for model training, and evaluates model performance with various metrics. The software introduces innovative model architectures and data augmentation for better results, presenting findings with clear visualizations. It provides a user-friendly interface (optional) and thorough documentation for ease of use. This software helps enhance early skin cancer detection, potentially saving lives.

#### a) Training Process

##### Sample python code for Training Process:

```
import torch
import torch.nn as nn
import torch.optim as optim
# Define your neural network model, loss function, and optimizer
model = YourModel()
criterion = nn.CrossEntropyLoss()
optimizer = optim.SGD(model.parameters(), lr=0.01)
# Number of training epochs
num_epochs = 100
for epoch in range(num_epochs):
    running_loss = 0.0
    for inputs, labels in dataloader: # Replace dataloader with your data loading mechanism
        optimizer.zero_grad()
        outputs = model(inputs)
        loss = criterion(outputs, labels)
        loss.backward()
        optimizer.step()
    running_loss += loss.item()
    average_loss = running_loss / len(dataloader)
    print(f'Epoch [{epoch + 1}/{num_epochs}] Loss: {average_loss:.4f}')
print("Training finished")
```

Now the forward propagation is done. The output is calculated. The loss is measured between the actual value and the target value. The model then computes the gradient of its parameters from the criterion. The parameters calculated are updated using

step 3.1. Then the whole training loss is calculated by adding up the values in each loss in each iteration. All the functions in the above code and the process involved in training are explained above.

The loss in each epoch is decreasing. The training dataset is divided into different batches as it will be difficult to train the whole batch of the training set at once and update the parameters based on the gradient descent for the whole training dataset. Different batches are trained at different epochs and the parameters based on the calculated gradient descent are updated for that particular batch only. Mini batch gradient descent is used for training. The gradient can further be reduced by adding the gradient over different batches.

**b) Testing Process**

A model is trained on the training dataset. It learns all the patterns and the way the model can classify the output based on the images. Based on this, the model that is trained will be able to test the testing images as it gets learned by the training process on the training dataset. The loss is calculated in the testing process. The parameters won't get updated. The best final parameters get updated in the training process. By using those parameters, an image is classified as benign or malignant. The image is sent to the first layer, and the loss is calculated based on the output value and the target value. The final loss is the sum of all the losses in each step. An image is tested on the model which is developed by the training.

**c) Classification**

Classification follows the testing process mentioned above. Using the Deep Learning Model developed, an image is sent as input to the developed model, and it is pre-processed. After that, the model calculates the loss for the image given based on the target value and the actual one. Finally, the model classifies the image into either a malignant or a benign form of cancer.

To implement the results obtained for your comparative study of machine learning models for early detection of skin cancer using Python, you can use libraries such as TensorFlow and Keras for developing and evaluating the models. Here's a Python script that demonstrates how to calculate and print the accuracy scores for CNN, VGG16, ResNet50, and AlexNet models, considering transfer learning.

```
print('Test Loss: %.6f\n'%(test_loss))

print('Test Accuracy: %2d%% (%2d/%2d)' % ((100. * correct/total),correct,total))

] model.cuda()

test(data_test,model,criterion,use_cuda)

Test Loss: 0.612456

Test Accuracy: 84% (422/500)
```

Fig 4. Accuracy for VGG16

- a) The test accuracy obtained for CNN for VGG16 is 84%.
- b) The test loss obtained is 0.61.

```
Model cuda()

Test(data_test,model,criterion,use_cuda)

Test loss:0.612456

Test Accuracy:81%(422/500);
```

Fig 5. Accuracy for AlexNet

- a) The test accuracy obtained for AlexNet is 81%.
- b) The test loss obtained is 0.61.
- a) The test accuracy obtained for ResNet is 82%.
- b) The test loss obtained is 0.61.

The metrics calculated are Accuracy, F1 score, and Classification report. The below image compares the accuracies of 4 different deep learning models developed.

$$\text{Accuracy} = ( \text{Correctly predicted class} / \text{Total testing class} ) \times 100\% \tag{2}$$

```

total += data.size(0)

print('Test Loss: %.6f\n'%(test_loss))

print('Test Accuracy: %2d%% (%2d/%2d)' % ((100. * correct/total),correct,total))

model.cuda()

test(data_test,model,criterion,use_cuda)

Test Loss: 0.612456

Test Accuracy: 82% (422/500)
    
```

Fig 6. Accuracy for ResNet

The comparison results from the existing study reveal the accuracy achieved by different models. The CNN model achieved an accuracy of 78%, while VGG16 outperformed with an accuracy of 84%, closely followed by ResNet50 with 82%, and AlexNet with 81%. VGG16 demonstrated the highest accuracy among these models, while the reference to "78%" appears to be incomplete or erroneous. It's important to clarify the accurate value. The models, other than the CNN, employed transfer learning, leveraging pre-trained layers, which significantly reduced training time for these transfer learning models.

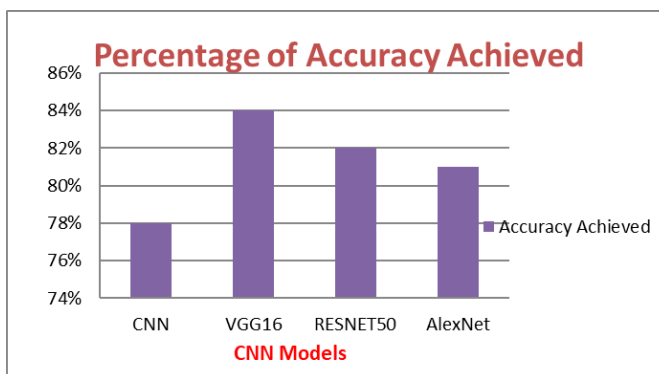


Fig 7. Accuracy-related comparison graph for CNN Models

The comparison results from the existing study reveal the accuracy achieved by different models. The CNN model achieved an accuracy of 78%, while VGG16 outperformed with an accuracy of 84%, closely followed by ResNet50 with 82%, and AlexNet with 81%. VGG16 demonstrated the highest accuracy among these models, while the reference to "78%" appears to be incomplete or erroneous. It's important to clarify the accurate value. The models, other than the CNN, employed transfer learning, leveraging pre-trained layers, which significantly reduced training time for these transfer learning models.

Figure 7 presents a comparison graph showcasing the accuracy of various algorithms. Notably, VGG16 exhibits the highest accuracy. VGG16 is particularly well-suited for images with relatively simple features, lacking substantial depth. This model excels in extracting straightforward characteristics like thickness, brightness, skin lesions, darkness, and skin color. VGG16's architecture, with a focus on Convolutional layers, contributes to its superior performance. It employs 3x3 filters with a stride of 1, and it consistently uses "same padding." Additionally, the network includes max pooling with 2x2 filters, and a stride of 2. In contrast, ResNet50, with its depth of 50 layers, excels in extracting more complex features from images due to its deeper architecture.

#### 4 Conclusion

The findings of this study underscore the critical importance of early identification of skin cancer, particularly melanoma. Notably, the research highlights substantial differences in the 5-year survival rates across various stages of melanoma, with stage 1 exhibiting a 90–95% survival rate and stage 4 suffering from a significantly lower 15-20% survival rate. The study demonstrates the potential of machine learning algorithms in effectively distinguishing between benign and malignant skin lesions in images, promising improved outcomes in terms of early detection and subsequent treatment.

Furthermore, this research introduces innovation by specifically concentrating on melanoma and integrating cutting-edge deep learning techniques into the pressing need for enhanced skin cancer diagnosis. Noteworthy contributions include the development of novel model architectures, the implementation of data augmentation techniques, and the introduction of innovative evaluation metrics. These innovations set this approach apart from existing methods and open up a fresh avenue for early skin cancer diagnosis. This underscores the value of continuous research and data collection in the critical realm of cancer detection, promising improved early diagnosis and ultimately more effective treatment strategies in the future.

## References

- 1) American Cancer Society. Key Statistics for Basal and Squamous Cell Skin Cancers. Website: American Cancer Society - Key Statistics for Basal and Squamous Cell Skin Cancers. American Cancer Society. 2021.
- 2) Skin Cancer Facts & Statistics. Website: Skin Cancer Foundation - Skin Cancer Facts & Statistics. 2021.
- 3) Kaur R, Gholamhosseini H, Sinha R. Deep Convolutional neural network for melanoma detection using dermoscopy images. *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. 2020;p. 1524–1527.
- 4) Sanketh RS, Bala MM, Reddy PVN, Kumar GVSP. Melanoma disease detection using Convolutional neural networks. *4th International Conference on Intelligent Computing and Control Systems (ICICCS)*. 2020;p. 1031–1037.
- 5) Hasan M, Barman SD, Islam S, Reza AW. Skin Cancer Detection Using Convolutional Neural Network. *Proceedings of the 2019 5th International Conference on Computing and Artificial Intelligence*. 2019;p. 254–258.
- 6) Naeem A, Farooq MS, Khelifi A, Abid A. Malignant melanoma classification using deep learning: datasets, performance measurements, challenges, and opportunities. *IEEE*. 2020;8:110575–110597. Available from: <https://doi.org/10.1109/ACCESS.2020.3001507>.
- 7) Rajasekhar KS, Babu TR. Skin Lesion Classification Using Convolution Neural Networks. *Indian Journal of Public Health Research & Development*. 2019;10(12):118.
- 8) Malo DC, Rahman MM, Mahbub J, Khan MM. Skin Cancer Detection using Convolutional Neural Network. *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*. 2022;p. 169–176. Available from: <https://doi.org/10.1109/CCWC54503.2022.9720751>.
- 9) Winkler JK, Schäfer I, Bender C. Diagnostic performance of a Convolutional neural network for dermoscopic melanoma recognition in clinical practice. *Journal of Investigative Dermatology*. 2021;141(11):2495–2498.
- 10) Brinker TJ, Hekler A, Enk AH. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*. 2019;111:148–154. Available from: <https://doi.org/10.1016/j.ejca.2019.04.001>.
- 11) Munir S, Chae M, Kim S. Dermatologist-level classification of melanoma with deep neural networks. *Journal of Investigative Dermatology*. 2021;141(11):2531–2534.
- 12) Sood S, Patel RS, Mangina S, Lee K. Deep learning for skin cancer detection using Convolutional neural networks. *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2019;p. 1846–1850.
- 13) Hekler A, Utikal JS, Enk AH. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer*. 2019;118:91–96. Available from: <https://doi.org/10.1016/j.ejca.2019.06.012>.
- 14) Smith JR, Johnson AB. A Comparative Study of Machine Learning Models for Early Detection of Skin Cancer Using Convolutional Neural Networks. *Proceedings of the 2021 International Conference on Machine Learning (ICML'21)*. 2021;p. 123–135.

# SUGAR CANE LEAF DISEASE CLASSIFICATION AND IDENTIFICATION USING DEEP MACHINE LEARNING ALGORITHMS

LAKSHMIKANTH PALETI<sup>1</sup>, ARAVA NAGASRI<sup>2</sup>, P. SUNITHA<sup>3</sup>, V SANDYA<sup>4</sup>, T. SUMALLIKA<sup>5</sup>, PRABHAKAR KANDUKURI<sup>6</sup>, K. KISHORE KUMAR<sup>7</sup>

<sup>1</sup>Associate Professor, Department of CSBS, R.V. R & J.C College of Engineering, Chowdavaram, A.P, India.

<sup>2</sup>Assistant Professor, Department of Computer Science and Engineering (Data Science), CVR College of Engineering, Ranga Reddy Dist., Telangana State, India.

<sup>3</sup>Associate Professor, Department of AI&DS, K L Deemed to be University, Green Fields, Vaddeswaram, Andhra Pradesh 522302, Email: <sup>4</sup>Assistant professor, Department of CSE (Data Science), CMR Technical Campus (Autonomous Engineering College), Kandlakoya, Telangana 501401

Email: <sup>5</sup>Assistant professor, Information Technology, Seshadri Rao Gudlavalleru Engineering College, Gudlavalleru

<sup>6</sup>Andhra Pradesh 521356. Email <sup>6</sup>Professor Department of AI & ML, Chaitanya Bharathi Institute of Technology, Osman Sagar Rd, Kokapet, Gandhi pet, Hyderabad, Telangana,

<sup>7</sup>Assistant professor, Department of CSE (Data Science), CMR Technical Campus (Autonomous Engineering College), Kandlakoya, Telangana 501401

Email: lakshmikanthpaleti@gmail.com, nagasri.arava@gmail.com, drsunitha.pachala@kluniversity.in, sandya.vooradi@gmail.com, sumallika.tella@gmail.com, prabhakarcs@gmail.com, kishorkadari@gmail.com

## ABSTRACT

The identification of crop diseases is one of the major concerns that the agricultural industry has to deal with. The detection and classification of leaves is essential in agriculture, forestry, rural medicine, and other commercial applications, among other things. The diagnosis of sugar cane plant leaf disease is required for automatic weed identification in precision agriculture. This paper discusses a novel approach to the development of a plant disease recognition model that is based on sugar cane leaf image classification and employs deep convolutional networks to recognise disease in sugar cane plants. The method used for identification and automatic recognition investigates the possibility of using k-NN and SVM in pre-training with ANN, followed by CNN-based approaches for recognition.

**Keywords:** *KNN, SVM, Leaf Disease, Classification, ML*

## 1. INTRODUCTION

Earlier those diseases were of minor importance but it has become matter of concern as it is sporadic rapidly in sugarcane growing area heavily monoculture of single variety and due to lack of effective technologies. But now a day's many techniques are applied this sector to predict or detect crops diseases. Some of the techniques are Image Processing, Machine Learning, Deep convolutional Neural Network (DCNN), Support Vector Machine (SVM), Public Neural Network

(PNN) [1]. Many researchers are publishing their paper to follow those techniques. Many scientists have already achieved a significant improvement in all those techniques. In this research I apply Deep Convolutional Neural Network (CNN) which is the advanced method of machine learning. The Sugarcane industry within the Bangladeshis contributes high profits to the economy. It is one of the biggest crops cultivated in several provinces round the country [2]. This crop provides 3 major products: sugar, bio-

ethanol, and power. At present, sugarcane is cultivated in regarding a hundred countries.

The cane industry produces approximately 5.5 million tonnes of cane per year. It is considered to be one of the most important money plants in the country. A sugarcane plant has stalks that are prominently jointed and bear two ranks of sword-shaped but gracefully arching leaves on each of their stalks. Some varieties may also have stalks that are between 5 and 7 metres in length. Sugarcane grows to its full potential in a tropical climate with rich, moist soil, bright sunlight, and warm temperatures

[3]. Among the best soils for sugarcane cultivation are clay-loam soils that contain a small amount of sand and silt and are rich in organic matter. Bangladesh's modern sugar manufacturing accounts for only about 5% of the country's total sugar consumption. Jiggery production, which is primarily based on sugarcane, accounts for approximately 20% of total demand, with the remaining 75% of total demand being met by imports. The primary reasons for the decrease in sugar production at the company include a decrease in the supply of sugarcane in the factories, which is due to the fact that the majority of the sugar is affected by one-of-a-kind diseases, and a decrease in the number of employees.

Diseases Sugarcane is susceptible to a number of diseases at various stages of its growth. All of these diseases are the most common in Bangladesh, and they are preventing the country from cultivating more sugarcane. Sugarcane crops are being destroyed at a rate of 30 to 40% per year because of this practise. So, in order to alleviate sugarcane diseases, we can employ some techniques that will produce a more favourable outcome. In this work, A popular technology at the moment is the use of machine learning to classify and detect plant diseases, and this is becoming increasingly common [4]. In order to use this method, more complex calculations must be performed, which can be time-consuming when using online applications. The performance of these methods, as a result, may only produce a satisfactory result in some instances.

Compared to the traditional architecture of the neural networks, profound learning uses artificial

neural network architecture which usually has many layers of information processing and is more sophisticated than regular neural network topologies. Deep learning has resulted in considerable increases in performance in the domains such as picture identification, image classification, acoustics and other sectors requiring extensive data processing. A profound learning for the detection of plant disease has prepared the road for the analysis and decision-making of professionals in the field [5]. The primary deep learning method in this study has been the Convolutionary Neural Networks (CNNs), which accounted for most of the findings. CNN technology is utilised as one of the most common approaches for demonstrating and relying on a big quantity of data, for pattern recognition applications.

Once the system has detected the image, we place some images for training and testing purposes, and then demonstrate the accuracy of the image result once it has been detected by system. A trend toward escapade in deep learning methodology for disease recognition is being observed as deep learning techniques advance and are applied in more and more applications in the following years [6]. There are several components to a CNN model depicted in Figure 1. These components include an input image, convolutional layer(s), pooling layer(s), fully connected layer(s), activation function(s), and an output. The components of a CNN model are depicted in Figure 1.

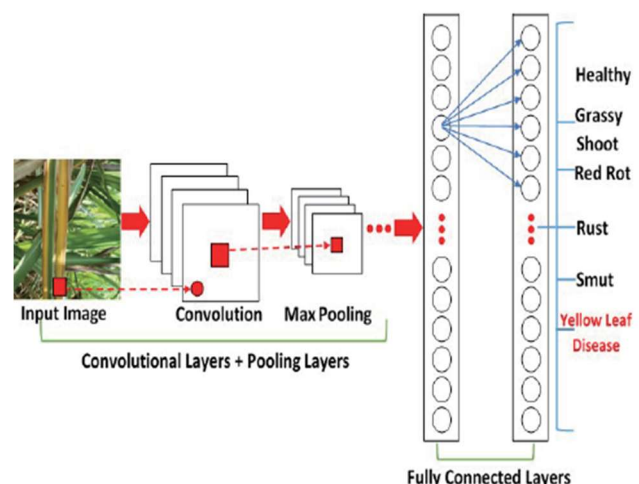


Fig. 1: Architecture For CNN



**2. RELATED WORK**

In the collection of disease images, four classes of diseases have been selected, with each class containing images of a different type of disease. Those all-class images are also assigned to a class, one of which is train data and the other which is testing data, and so on [7]. All of the photographs were taken with a mobile phone on sugarcane land. When collecting images, make an effort to find images of high quality. The majority of the images are not disordered; some images are also in good health. Red rot, sugarcane borer, rust, and wilt are the diseases that affect sugarcane.

**Red Rot**



*Fig 2: Image For Red Rot Sugar Cane Disease*

**Sugarcane Borer**



*Fig 3: Image For Sugarcane Borer Disease*

**Rust**



*Fig 4: Image For Rust Sugar Cane Disease*

**Wilt**



*Fig 5: Image For Wilt Sugar Cane Disease*

**a. Classification**

Deep learning was effectively used in a number of applications such as the detection and classification of crop varieties, plant identification and classification, picture grading of fruits and vegetables, and image classification. A rise in popularity has also been seen in photographs taken with mobile cameras, as well as photographs taken with any camera device mounted on a robot. Convolutional Neural Networks, also known as CNNs, are becoming increasingly popular among computer vision researchers, particularly in the field of computer vision, due to their ability to execute different layers of processing through multiple stages of execution. Because of this, CNNs are becoming increasingly popular among computer vision researchers, particularly in the field of computer

vision [8]. Using an architecture of convolutional neural networks to illustrate the process, Figure 6 depicts several stages in the prediction of plant disease at various stages in the process. Using an architecture of convolutional neural networks, the figure 6 depicts several stages in the prediction of plant disease. It is proposed that the proposed work include a detailed description of how the model will be put into practise.

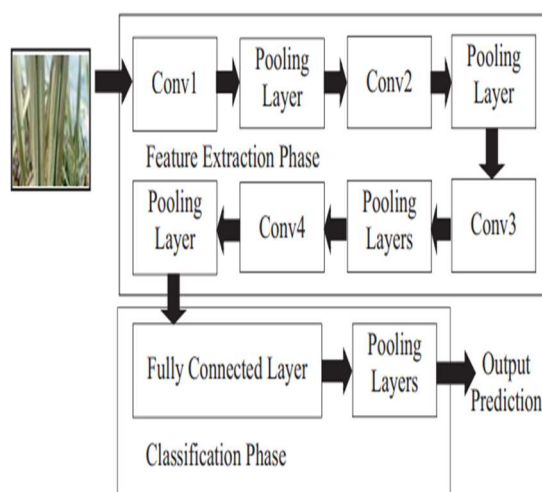


Fig. 6: Disease Prediction Using CNN Architecture

### b. Detection process

In order to determine how well two cutting-edge detection networks, YOLOv3 and Faster-RCNN, perform when it comes to identifying infected regions in images, they will be tested and evaluated in this experiment. Models such as R-CNN and Fast R-CNN have been developed in the past, but the aforementioned models are significantly more rapid than those that came before them. When it came to finding the regions on which CNN will be passed separately for classifying the label, it used a more time-consuming method known as selective search, which was significantly more time-consuming than the previous method [9]. In order to generate a small number of thousand regions of interest, each one was generated and passed separately to the network for further classification and analysis. This method was created with the intention of being unsuitable for real-time inference. In the case of Faster-RCNN, a region proposal network

(RPN) is used to predict region proposals on a convolutional feature map after the image has been passed through a CNN, as opposed to a CNN alone.

We trained the complete model on ImageNet dataset, starting with pretrained block weights in the two models and on from there, to assess the realisation of both architectures in our dataset. The quicker R-CNN was trained and evaluated using images with the same resolution at a resolution of 600x1000 pixels across 15 epochs. It was trained on images with a resolution of 416x416 for 6000 iterations before being tested on images with resolutions of 416x416 and 608x608 for a total of 6000 iterations on 416x416 and 608x608 images.

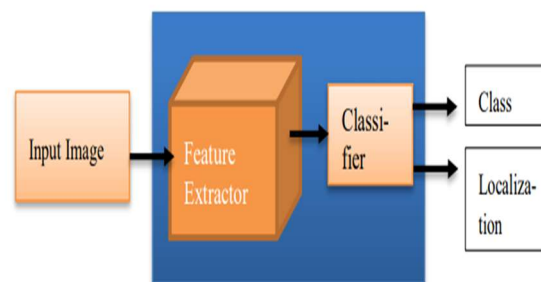


Fig. 7. Pictorial Representation Of Detection Process.

### 3. LIMITATIONS OF THE EXISTING WORK

In organic farming, crop protection is a tricky subject. It demands a thorough understanding of the crops being farmed, as well as any pests, illnesses, or weeds present. Based on particular convolutional neural network designs, our system created specialised deep learning models for identifying plant illnesses using leaf pictures of healthy or diseased plants. Our detector combined photos from a variety of sources with photos collected on-site by various camera systems. The algorithms used in this paper and their Limitations; advantages are listed in the below table.

#### PROBLEM STATEMENT

This paper addresses the critical need for automated Sugar Cane Leaf Disease Classification and Identification using Deep

Machine Learning Algorithms. Sugar cane crop health directly impacts yield and quality, and the project's main objectives include collecting and preprocessing a diverse dataset of sugar cane leaf images, selecting appropriate deep learning models like Convolutional Neural Networks (CNNs), implementing feature extraction and pattern recognition techniques, rigorous training and validation, and the development of a user-friendly interface for instant disease diagnosis. Performance evaluation metrics will be used to continually improve the model, ensuring scalability to accommodate a growing user base, ultimately aiding sugar cane farmers and agricultural experts in prompt disease detection and crop management for enhanced yield and sustainability.

		set contains more noise, such as overlapping target classes, SVM does not perform well. The SVM will underperform if the number of features for each data point exceeds the number of training data samples.	is a clear separation between classes. SVM is stronger in high-dimensional spaces. SVM becomes effective when the number of dimensions exceeds the number of samples. The SVM algorithm was created with memory conservation in mind.
3	ANN	Operation of the Network That Isn't Clearly Explained System Requirements Make sure the network's structure is correct. The difficulty of informing the network about the situation The network's lifespan is unknown.	ANNs have several major advantages that make them ideal for a variety of issues and scenarios: ANNs can learn and represent non-linear and complicated interactions because many of the relationships between inputs and outputs in real life are both non-linear and intricate.
4	CNN	The position and orientation of an object are not	CNN has a significant advantage

Sn o	Meth od	Limitations/De merits	Merits
1	KNN	<ul style="list-style-type: none"> <li>The quality of the data determines its accuracy.</li> <li>The prediction stage may take a long time if there is a lot of data.</li> <li>Aware of the data's size and irrelevant characteristics.</li> <li>Requires a large amount of memory due to the fact that all of the training data must be saved.</li> <li>Because it stores all of the training, it can be computationally expensive.</li> </ul>	<p>Calculation time is limited.</p> <p>To decipher a straightforward algorithm. It has a wide range of applications, including regression and classification.</p> <p>There's no need to compare to more supervised learning models because of the high precision.</p>
2	SVM	For big data sets, the SVM algorithm is ineffective. When the data	SVM performs reasonably effectively when there

		<p>encoded by CNN. A convolutional layer is the most important part of a CNN. Inability to be spatially invariant when dealing with incoming data. A single scalar is produced by artificial neurons. What's the best way to cope with CNN?</p>	<p>over its predecessor s in that it can detect crucial characteristics without the requirement for human interaction. It can learn distinctive features for each class on its own given a sufficient number of images of cats and dogs. Furthermore, CNN is a computationally efficient algorithm.</p>
--	--	---	---

using computer software. In order to differentiate between diseased and healthy images of sugarcane leaves, each image is saved in its own folder with a label indicating which class it belongs to. 3295 images were collected and organised into seven different categories in the image dataset that was acquired. Each image is saved in the uncompressed JPG or PNG format, and it is coloured using the RGB colour space as a base colour.

**B. Pre-processing of Images**

Pre-processed images include images that have been reduced in size, images that have been cropped, and images that have been enhanced. For the purposes of this study, we have used coloured images that have been resized to a resolution of 96x96 pixels in order to be processed further.

**C. Feature Extraction**

The convolutionary layers extract characteristics from scaled images. The nonlinear activation function Rectified (ReLU) is applied after convolution with the purposes of reducing the size of features extracted, using various methods of packing such as maximum pooling and average pooling. When the convolution and pooling layers are combined, the result will be a filter that generates features for analysis.

**D. Classification**

Classification is accomplished through the use of fully connected layers, and feature extraction is accomplished through the use of convolutional and pooling layers.

Method	Normal Accuracy prediction	Training parameters	Accuracy prediction by cross validation
KNN	75	K=13	75.02
SVM	93	Rbf Kernel	93
ANN	61	5-15 Neurons	26
CNN	88	0-288	87

**4. METHODOLOGY**

Figure 8 illustrates a process diagram based on experimental design that shows if the sugarcane plant is infected or not with leaf pictures from the disease [10].

**A. Image Dataset Acquisition**

Images of sugarcane leaves are captured manually with a camera and then enhanced and segmented

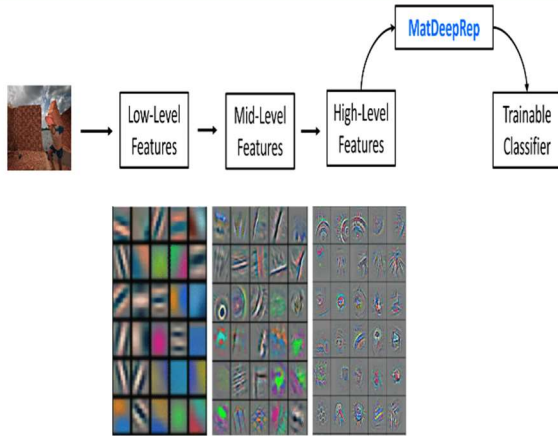


Figure-7.1 Classification

This layer is responsible for classifying the sugarcane leaves and determining whether or not they are infected with the disease.

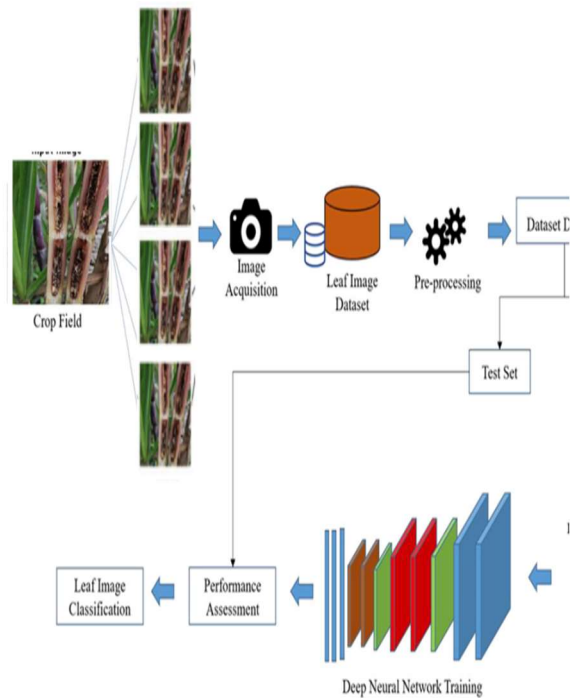


Fig 8: Working Of The Classification Of The Sugarcane Leaf Disease

The entire procedure of creating a replica for plant disease identification through the use of deep CNN is described in detail right here in this document. The entire system is divided into a few

critical steps, starting with the recruitment of images for the classification system and progressing to the application of deep neural networks. Figure 9 depicts an experimental design based on a workflow diagram that, through the use of images, indicates whether or not the sugarcane plant has been infected with the disease, as well as the results of the experiment [11]. Figure 9 Experimental design based on a workflow diagram.

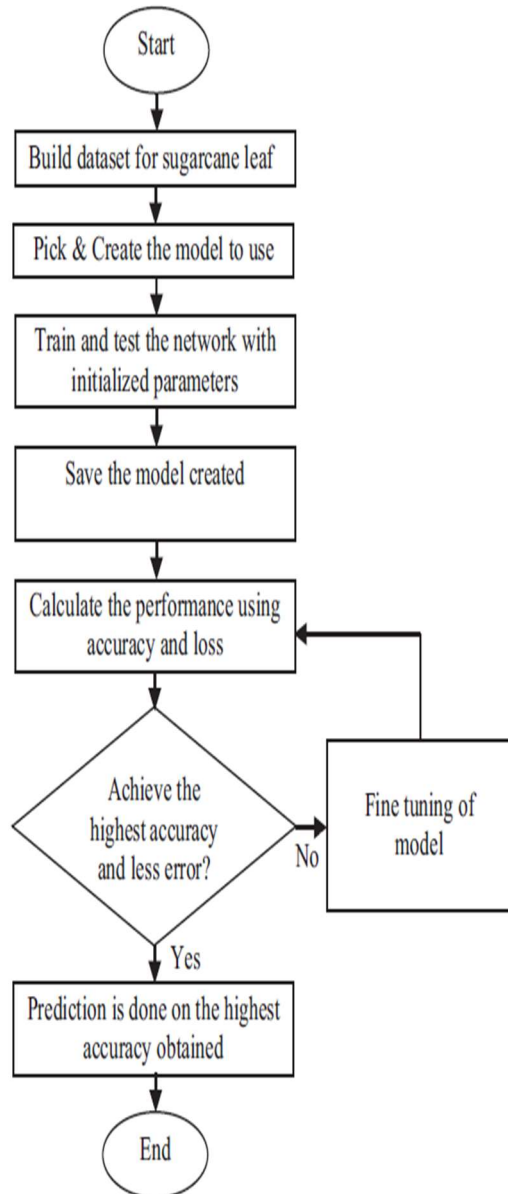


Fig. 9: Prediction Model Flow Chart

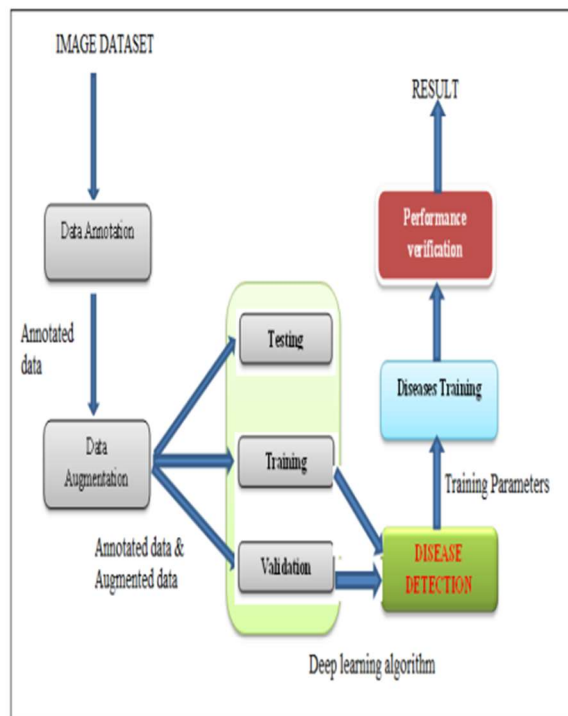


Fig 10: Working Of Deep Learning Algorithm

of the photos in the data set have several diseased spots and they have various patterns. The different parks in accordance with their geographical position were noted individually for all these places. Table 1 offers further information on each category, which illustrates also how the photographs have been sorted into distinct categories. Figure 11 shows the classes utilised for classification and detection and their relative distributions and distributions.

Table 1: The Distribution Of The Images Into Different Classes

S. No	Classes	Count
1	Red Rot	545
2	Rust	832
3	Sugarcane Borer	570
4	Wilt	420
5	Healthy	928

### Data Set

The dataset contains 3295 images of sugarcane leaves, which are divided into six different categories by their shape and size (consisting of 4 diseases and 1 healthy). It is these diseases that have caused the most serious damage to Indian crops over recent years that are listed here [12]. All of the photographs were taken in a natural setting with a wide range of variations in light and composition.

These images were taken from a range of cultivation areas, including those at Mandya Bangalore's Agricultural Science University and farms owned by farmers in the vicinity. Everything in a range of ways, orientations and backdrops was shot using telephone cameras and thus represents the vast majority of changes that can appear in real world images. The sample collection was assisted by a group of pathologists who are well knowledgeable in their profession. We manually annotated the dataset for tracing the sick spots on the leaves (object detection) that match four different diseases [13]. The majority

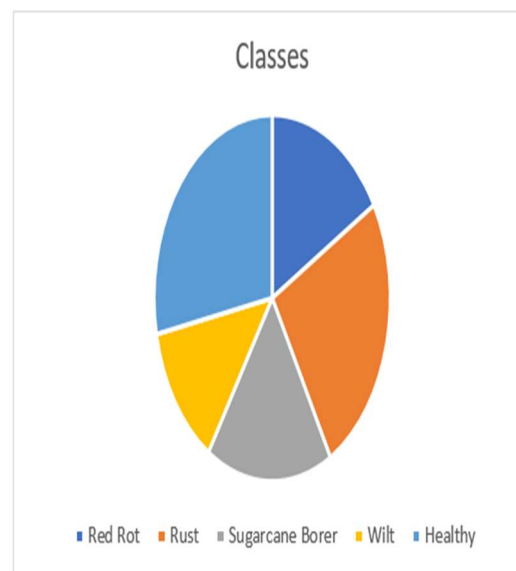


Figure 11: The Distribution Of The Images Into Different Classes

The test is a critical component of this research in order to achieve higher overall performance. We have already trained the data set using a convolutional network process, and the information about the training is available to us at this point [14]. Figure 12 shows how to select a train image as input, after which the image is sent to exhaustive search, which aids in the images enumerating all possible tasks, and then it is sent to CNN for the purpose of producing an output image [15]. The output image is checked by the SoftMax classifier before being used. When we talk about SoftMax classifier, we are talking about a loss function, which in the context of Machine Learning and Deep Learning tells us to quantify how good or bad a given classification term is at precisely classifying data points in our data set.

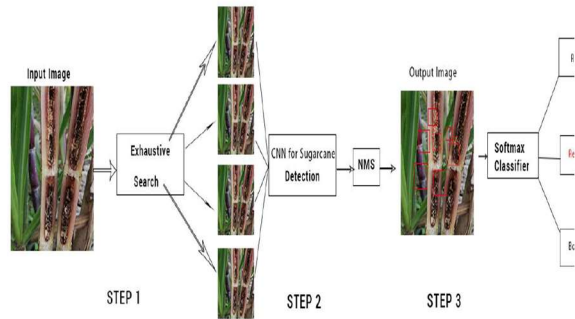


Fig 12: Overall Testing Process

## 5. RESULTS

The overall process graph is depicted in the preceding figure 13, which is shown in the preceding paragraph. For the purposes of this chapter, we will set up all of the images with various classes in an experimental manner. After that, the images are rotated to a 25-degree angle for enhancement, after which they are flipped and shifted horizontally and vertically to achieve the desired effect. When the batch size is set to 10, the model train will run for a total of 60 iterations. In the end, the Deep Convolutional Neural Network with Confusion Matrix brings everything together and provides the final result. It is possible to calculate the accuracy and error rate of a calculation by using a confusion matrix. Table 1 of the confusion matrix is shown in the preceding image. Figure 13 shows that the training accuracy is very close to the validation accuracy, which is

a good thing. Because of this method's use, we can say that training accuracy has been improved. Sixty-two epochs were used in the development of the training model, resulting in accuracy rates as high as 88 percent. The outcome was described as more favourable by the authors of the paper. Any solution that has an accuracy lower than 60% cannot be considered satisfactory. Figure 13 illustrates the recognition of plots of train and test accuracy when testing random images of sugarcane plant diseases. Figure 13 depicts the recognition of plots of train and test accuracy when testing random images of sugarcane plant diseases in addition, we can see the graphs of training loss and validation loss, which show that the training loss is decreasing slowly with each passing day as time progresses. In the example above, the training loss and validation loss are shown, and they are obtained after 60 epochs. It's simple to calculate the error rate from accuracy after you've finished the process, and an error rate of 8 percent indicates a more favourable outcome. Finally, we can state that the Convolutional Neural Network produces a better result and greater accuracy throughout the process.

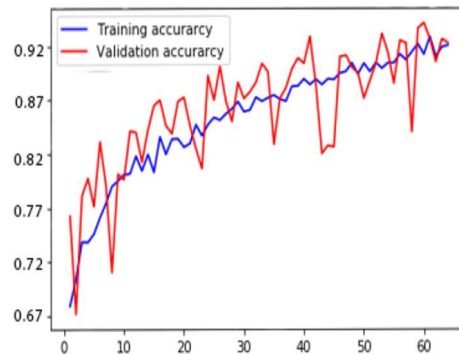
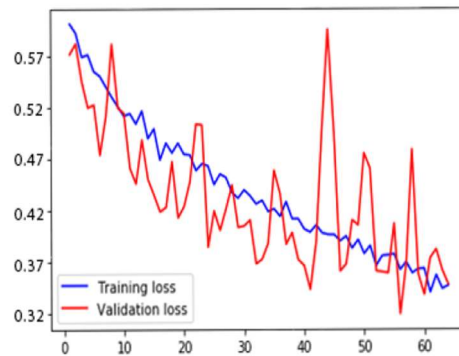


Table-3 Confusion Matrix

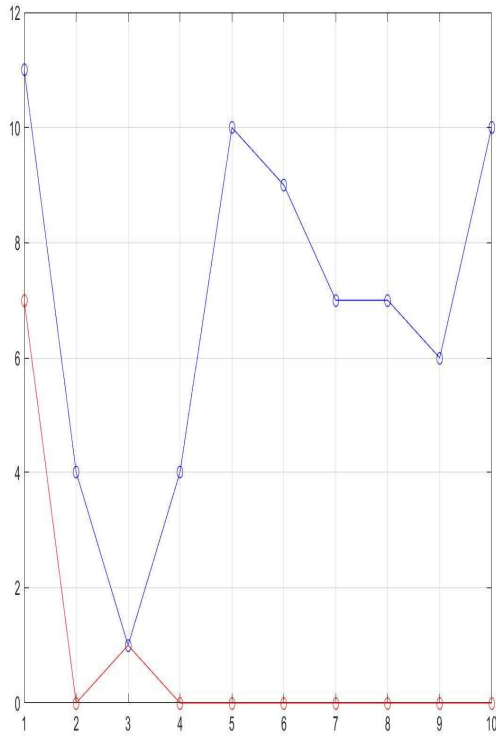


Figure 13: Training Accuracy, Validation Accuracy And Training Loss, Valadon Loss.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / \text{Actual No. of samples}$$

$$\text{Accuracy} = (397 + 2504) / 3295$$

$$\text{Accuracy} = 88\%$$

$$\text{Error rate} = 100 - 88 = 12\%$$

Table 2: Confusion Matrix

Sugarcane Diseases		Predicted class	
		Healthy	Infected
Actual Class	Healthy	397 (TN)	252 (FP)
	Infected	140 (FN)	2504 (TP)

1.	2	0.	6	0.	0.	3.			
6	0	0.	8	6.	1	7	8	0.	0.
7	8	8	0	6	3	1	5	8	0
6	2.	2	2	6	3	6	4	2	1
0	2	3	5	0	7	3	6	3	0
6	6	8	1	3	0	1	0	8	6
3	1	9	2	8	3	8	2	9	4
3.	1	0.	0.	7	0.	0.		0.	0.
5	1	6	9	3.	0	4		6	0
9	9	6	6	8	3	3	5.	6	0
2	2.	1	0	5	3	3	8	1	2
9	3	8	4	4	4	4	2	8	6
5	3	2	8	9	8	8	1	2	6
4	4	6	8	8	9	6	5	6	5
1.	1	0.	0.	5	0.	0.	4.	0.	0.
5	6	7	7	6.	0	7	7	7	0
1	8	1	5	9	7	0	7	1	0
3	3.	5	0	5	7	5	9	5	6
1	8	5	5	8	8	0	0	5	1
9	5	2	1	0	6	2	1	2	9
4	6	9	5	1	9	2	8	9	7
1.	3	0.	0.	2	0.	0.		0.	0.
6	2	8	7	5.	2	7	2.	8	0
0	7.	2	8	8	4	2	8	2	1
5	0	9	2	5	0	4	6	9	9
3	3	3	2	4	6	8	0	3	1
4	2	8	8	4	1	1	3	8	4
6	2	4	7	1	5	5	5	4	7
1.		0.	0.	4	0.	0.	6.	0.	0.
3	1	4	6	8.	0	6	7	4	0
8	3	4	9	7	3	6	2	4	0
8	4	9	3	1	2	1	9	9	2
6	2.	8	8	6	2	5	2	8	5
5	3	0	4	7	4	9	4	0	6
1	1	7	6	2	3	7	6	7	6
1.		0.	0.	1	0.	0.	7.	0.	0.
6	1	4	7	5	0	5	0	4	0
0	1	7	8	1.	2	1	3	7	0
6	2	5	2	4	0	4	0	5	1
7	0	0	7	0	0	3	8	0	5
0	4.	8	0	0	6	3	9	8	9
8	9	8	7	6	6	7	2	8	7
1.	9	0.	0.	4	0.	0.	3.	0.	0.
7	7	6	8	7.	1	6	3	6	0
9	0.	7	2	0	4	3	3	7	1





ISSN: 1992-8645

[www.jatit.org](http://www.jatit.org)

E-ISSN: 1817-3195

0	6	8	9	4	8	7	1	8	1	3.		0.	0.	6	0.	0.	3.	0.	0.	
4	9	1	4	1	7	3	1	1	8	9		7	9	5.	1	4	1	7	0	
7	1	1	9	3	5	7	7	1	3	5	8	4	6	4	1	2	9	4	0	
5	9	2	7	7	4	1	6	2	7	4	5	5	7	3	1	5	4	5	8	
2.	5	0.	0.	1	0.	0.	2.	0.	0.	2	0.	0	4	6	4	2	0	0	8	
2	4	8	8	2	3	6	2	8	0	3	4	1	9	2	3	2	1	1	6	
4	7	7	9	5.	8	4	6	7	3	8	8	6	4	6	9	6	4	6	8	
9	1.	5	5	1	1	3	6	5	0	2.	5	0.	0.	1	0.	0.	3.	0.	0.	
7	1	2	7	8	6	8	5	2	3	5	5	9	9	3	1	6	6	9	0	
9	7	1	8	9	5	2	4	1	7	1	9	2	1	3.	3	0	3	2	1	
1	6	3	6	1	8	2	9	3	1	4	1.	8	7	7	8	1	6	8	1	
3.	3	0.	0.		0.	0.		0.	0.	1	9	7	4	9	3	8	7	7	0	
8	2	6	9	1	0	4	4.	6	0	3	7	8	9	2	4	2	7	8	0	
4	1	8	6	2	7	4	1	8	0	7	8	4	4	7	1	8	7	4	9	
3	8.	4	5	5.	2	4	1	4	5	2.	3	0.	0.	9	0.	0.	3.	0.	0.	
4	7	1	5	5	7	9	3	1	7	1	4	4	8	6.	0	4	5	4	0	
2	4	9	5	0	0	2	5	9	8	3	0	8	8	1	9	9	2	8	0	
1	7	9	9	4	7	2	1	9	6	3	3.	9	3	3	0	6	6	9	7	
4.	1	0.	0.	8		0.	2.	0.	0.	0	1	0	2	7	8	8	4	0	2	
6	1	8	9	3.	0.	4	9	8	0	3	2	8	9	6	0	4	7	8	2	
8	7	0	7	6	1	3	2	0	1	8	7	5	6	4	9	7	1	5	6	
6	1.	8	6	0	4	2	6	8	1	1.	9		0.	1	0.	0.	5.		0.	
1	5	5	9	6	6	6	1	5	6	1	4	0.	5	1	0	5	5	0.	0	
2	4	4	6	7	5	9	8	4	6	8	4	3	4	9.	2	5	9	3	0	
2	4	2	6	3	7	1	8	2	4	9	0.	4	1	5	8	4	9	4	2	
6.	1	0.			0.	0.	1.	0.	0.	0	0	1	0	4	6	8	0	1	2	
6	0	9	0.	9	3	3	7	9	0	8	6	3	5	9	0	8	1	3	7	
2	7	7	9	5.	5	7	2	7	2	1	7	1	7	6	1	6	2	1	6	
9	3.	1	8	1	4	6	6	1	8	1.			0.	7	0.	0.	6.		0.	
9	1	0	8	7	0	2	1	0	1	4	3	0.	7	5.	0	5	1	0.	0	
3	6	6	5	9	7	6	5	6	7	9	2	1	3	1	3	2	4	6	3	0
2	3	6	6	3	9	4	9	6	7	5	2	4	2	2	6	2	8	4	2	
1.	1	0.	0.	1		0.		0.	0.	6	5.	6	7	3	4	7	0	6	1	
5	6	8	7	7	0.	7	4.	8	0	3	6	0	2	0	0	8	6	0	0	
1	0	7	5	6.	1	6	1	7	0	2	3	7	6	1	1	1	2	7	1	
9	3	1	2	1	2	7	4	1	9	8.	2	0.	0.	5	0.	0.	2.	0.		
3	1.	0	8	0	3	0	1	0	8	3	5	6	9	1.	1	3	4	6	0.	
8	4	4	7	6	1	7	0	4	0	6	1.	8	9	7	2	0	3	8	0	
8	5	8	8	7	7	8	5	8	2	6	3	5	2	4	2	3	0	5	0	
1	2		0.	6	0.	0.	1.		0.	2	6	6	8	5	5	1	1	6	9	
2.	7	0.	9	5.	1	2	8	0.	0	5	7	8	3	7	1	2	8	8	7	
5	1.	9	9	7	5	5	9	9	1	1	1	6	1	5	7	5	1	6	5	
0	2	2	6	2	4	6	7	2	2	8.	2	0.	0.	1	0.	0.			0.	
8	5	5	7	7	2	1	7	5	2	6	0	0.	7	1	0	4	5.	0.	0	
5	3	2	9	4	2	8	2	2	7	2	8	4	8	2.	3	8	3	4	0	
8	8	8	9	8	2	8	2	8	3	5	2.	1	8	2	1	3	0	1	2	
										9	5	9	5	1	4	1	4	9	4	

5	7	3	0	5		6	0	3	9
5	8	7	9	6		8	5	7	9
1.			0.	1	0.	0.	1.		0.
0			2	7	8	9	6		0
2	2		2	0.	2	6	8		6
6	2		7	3	9	4	5		6
8	2		1	7	6	3	1		0
4	0		6	0	0	4	2		1
6	1	1	5	3	4	7	5	1	8
1.	4		0.	2	0.	0.	4.		0.
0	5	0.	1	4	1	9	2	0.	0
2	1	9	9	2.	2	4	1	9	1
0	3	1	7	1	6	5	2	1	0
0	2.	6	2	0	7	4	1	6	0
3	6	9	3	7	8	3	4	9	8
8	2	8	9	7	6	6	1	8	9
1.	2	0.	0.	2	0.	0.	5.	0.	0.
3	4	7	6	0	0	8	5	7	0
4	1	4	6	3.	6	0	5	4	0
0	8	3	5	1	8	6	6	3	5
4	8.	7	9	8	1	3	5	7	4
4	1	3	2	0	7	1	4	3	2
9	6	2	5	3	1	5	4	2	5
3.	6	0.	0.	1	0.	0.	2.	0.	0.
3	5	9	9	6	2	5	7	9	0
5	1	6	5	6.	1	3	6	6	1
9	3.	7	4	9	6	2	8	7	7
2	1	9	6	0	5	2	7	9	2
1	2	6	6	4	8	5	6	6	3
6	6	9	3	8	5	7	6	9	5
	7	0.		1	0.	0.	6.	0.	0.
1.	2	3	0.	1	0	5	7	3	0
5	7	9	7	9.	2	1	9	9	0
5	7.	5	6	8	0	1	7	5	1
0	9	3	4	6	6	6	6	3	6
5	2	1	2	8	6	0	9	1	4
8	3	3	5	7	1	3	2	3	4

photographs of sugar cane into good and ill class, the trained model has made its intention mainly through leaves and stem samples. In future, a mobile application has been implemented on the basis of our search to detect the sugar cane leaves and stems disorder and to provide records of this disease. Artificial neural community (ANN) that we are able to easily detect plant damage items provided on Android phones. The Future work of this paper is to implement the system with below mentioned algorithms Faster Region-based Convolutional Neural Network (Faster R-CNN), Region-based Fully Convolutional Network (R-FCN), and Single Shot Multibox Detector (SSD),

**REFERENCES**

[1] S. K. Gupta and A. P. Agarwal, "Predicting Total Sugar Production Using Multivariable Linear Regression," 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2021, pp. 465-469, doi: 10.1109/ICCCIS51004.2021.9397078.

[2] R. Ekawati, Y. Arkeman, S. Suprihatin and T. C. Sunarti, "Design of Intelligent Decision Support System for Sugar Cane Supply Chains Based on Blockchain Technology," 2020 2nd International Conference on Industrial Electrical and Electronics (ICIEE), 2020, pp. 153-157, doi: 10.1109/ICIEE49813.2020.9276755.

[3] C. Hortinela et al., "Classification of Cane Sugar Based on Physical Characteristics Using SVM," 2019 IEEE 11th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management (HNICEM), 2019, pp. 1-5, doi: 10.1109/HNICEM48295.2019.9072699.

[4] Sammy V. Militante, Bobby D. Gerardo, Ruji M. Medina, "Sugarcane Disease Recognition using Deep Learning", Published 2019 Computer Science 2019 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)

[5] S. Militante, Fruit Grading of Garcinia Binucao (Batuan) using Image Processing, International Journal of Recent Technology and Engineering (IJRTE), vol. 8 issue 2, pp. 1829- 1832, 2019

**6. CONCLUSION**

This paper has been thoroughly trained in whether sugarcane leaves and stem are diseased or healthy by the Convolutional Neural Network (DCNN). The structure utilised to categorise the sugar cane leaf using a simple convolutionary neural community with 6 unique instructions, the accuracy achieved is 88% and the error rate of 12%. In order to efficiently detect and classify

- 
- [6] K. P. Ferentinos, Deep learning models for plant disease detection and diagnosis, *Computer Electronics Agriculture*, vol. 145, no. September 2017, pp. 311–318, 2018.
- [7] R. F. Rahmat, D. Gunawan, S. Faza, K. Ginting and E. B. Nababan, "Early Identification of Leaf Stain Disease in Sugar Cane Plants Using Speeded-Up Method Robust Features," 2018 Third International Conference on Informatics and Computing (ICIC), 2018, pp. 1-6, doi: 10.1109/IAC.2018.8780482.
- [8] A. Kamilaris and F. X. Prenafeta-Boldú, Deep learning in agriculture: A survey, *Computer Electronics Agriculture*, vol. 147, no. July 2017, pp. 70–90, 2018.
- [9] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, *MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications*, 2017.
- [10] K. P. Ferentinos, Deep learning models for plant disease detection and diagnosis, *Computer Electronics Agriculture*, vol. 145, no. September 2017, pp. 311–318, 2018.
- [11] H. Park, J. S. Eun and S. H. Kim, Image-based disease diagnosing and predicting of the crops through the deep learning mechanism, In *Information and Communication Technology Convergence (ICTC)*, IEEE 2017 International Conference on, pp. 129-131, 2017.
- [12] K. Elangovan and S. Nalini, Plant disease classification using image segmentation and SVM techniques, *International Journal of Computational Intelligence Research*, vol. 13(7), pp. 1821-1828, 2017.
- [13] A.K. Mahlein, Plant disease detection by imaging sensors-parallels and specific demands for precision agriculture and plant phenotyping, *Plant Disease*, vol. 100, no. 2, pp. 241-251, 2016.
- [14] S. P. Mohanty, D. P. Hughes, and M. Salathe Using Deep Learning for Image-Based Plant Disease Detection, *Frontier Plant Science*, vol. 7, no. September, pp. 1–10, 2016

# Improving Phishing Website Detection with Machine Learning: Revealing Hidden Patterns for Better Accuracy

Garlapati Narayana<sup>1</sup>, Uma Devi Manchala<sup>2</sup>, Usikela Naresh<sup>3</sup>, Saggurthi Kiran<sup>4</sup>, Medikonda Asha Kiran<sup>5</sup>, Ravi Kumar Ch<sup>6</sup>

<sup>1</sup>Associate Professor, Department of CSE (AIML),  
Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, Telangana, India  
narayanag.1973@gmail.com

ORCID: <https://orcid.org/0000-0001-8470-3595>

<sup>2</sup>Assistant Professor, Department of Computer Science and Engineering,  
Nalla Narsimha Reddy Education society's Group of Institutions, Chowdariguda, Medchal, Telangana, India  
umadevi.manchala92@gmail.com

ORCID: <https://orcid.org/0000-0002-8325-3868>

<sup>3</sup>Assistant Professor, Department of Computer Science and Engineering (AI&ML),  
CVR College of Engineering, Mangalpalli, Rangareddy, Telangana, India  
usikelanaresh@gmail.com

ORCID: <https://orcid.org/0009-0006-9656-4880>

<sup>4</sup>Assistant Professor, Department of Computer Science and Engineering (AI&ML),  
CMR Technical Campus, kandlakoya, Medchal, Telangana, India  
kiransaggurthi@cmr.edu.in

ORCID: <https://orcid.org/0009-0002-5997-2288>

<sup>5</sup>Assistant Professor, Department of AIML,  
Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, Telangana, India  
ashakiran2@gmail.com

ORCID: <https://orcid.org/0000-0002-7760-2902>

<sup>6</sup>Assistant Professor, Department of AI&DS,  
Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad, Telangana, India  
chrk5814@gmail.com

ORCID: <https://orcid.org/0000-0003-0809-5545>

**Abstract:** Phishing attacks remain a significant threat to internet users globally, leading to substantial financial losses and compromising personal information. This research study investigates various machine learning models for detecting phishing websites, with a primary focus on achieving high accuracy. After an extensive analysis, the Random Forest Classifier emerged as the most suitable choice for this task. Our methodology leveraged machine learning techniques to uncover subtle patterns and relationships in the data, going beyond traditional URL and content-based restrictions. By incorporating diverse website features, including URL and derived attributes, Page source code-based features, HTML JavaScript-based features, and Domain-based features, we achieved impressive results. The proposed approach effectively classified the majority of websites, demonstrating the efficiency of machine learning in addressing the phishing website detection challenge with an accuracy of over 98%, recall exceeding 98%, and a false positive rate of less than 4%. This research offers valuable insights to the field of cyber security, providing internet users with improved protection against phishing attempts.

**Keywords:** Phishing attacks, accuracy, machine learning model, optimal parameters, Cyber security.

## I. INTRODUCTION

The internet has revolutionized the way we conduct business, communicate, and access information. However, this digital transformation has brought about a dark side: cybercrime. Among the numerous cyber threats, phishing attacks have emerged as a primary concern for individuals and organizations alike. Phishers employ social engineering techniques to manipulate human vulnerability, luring

unsuspecting victims into revealing sensitive information or performing actions that can have dire consequences [1][2].

Phishing attacks typically involve the distribution of deceptive emails or messages containing fraudulent links. Once recipients fall into the trap, cybercriminals exploit this opportunity to gain unauthorized access to victims' accounts, leading to financial loss, identity theft, and other severe ramifications. Despite efforts to mitigate this menace, the

proliferation of phishing websites and the evolution of sophisticated tactics have made traditional detection methods less effective [3].

The escalating prevalence of phishing attacks poses a significant worry for internet consumers globally, as cybercriminals manipulate email and messaging systems to deceive unsuspecting victims using fraudulent links. Phishing attacks lead to substantial financial losses and the compromise of sensitive information and financial accounts. Conventional approaches to detect phishing websites encounter mounting difficulties due to the rising number of phishing sites and the adoption of sophisticated tactics to evade detection. This literature review examines previous research on machine learning-based methodologies to enhance the identification of phishing websites, aiming to tackle these challenges and protect internet users from the pervasive threat of cybercrime [4].

### **1.1 Challenges with Traditional Methods**

Traditional approaches for detecting phishing websites have long relied on techniques like visual verification, content-based analysis, and maintaining blacklists of known phishing URLs. Although effective in the past, these methods struggle to keep pace with the ever-increasing number of phishing sites and the cunning techniques employed by phishers. Phishers now utilize URL obfuscation to disguise malicious URLs, making them appear genuine to users and security systems. Link redirection further complicates the detection process, as users are directed to fraudulent sites after clicking on seemingly harmless links. Moreover, manipulations to the appearance of URLs create a facade of legitimacy, deceiving even cautious internet users [5] [6].

### **1.2 The Machine Learning-Based Approach**

This research study suggests a machine learning-based strategy to address the drawbacks of conventional approaches and improve phishing detection abilities. Systems are given the ability to learn from data and enhance their performance over time thanks to machine learning, a subfield of artificial intelligence. Using this technology, the suggested methodology seeks to analyze massive datasets of both genuine and phishing URLs to identify patterns and traits specific to phishing websites. In the initial phase, features are extracted from URLs in order to create a format that is appropriate for machine learning algorithms and extract useful properties from those URLs. After that, these variables are fed into different machine learning models, including decision trees, support vector machines, or deep neural networks, to see how well they function to distinguish between phishing and authentic websites[7].

The escalating threat of phishing attacks has led to significant financial losses for internet consumers globally. Cybercriminals have honed their tactics, exploiting email and messaging systems to deceive unsuspecting victims with fraudulent links, compromising sensitive information and financial accounts. Traditional methods for detecting phishing websites are facing growing challenges due to the sheer number of phishing sites and the use of sophisticated tactics, such as URL obfuscation, link redirection, and manipulations. To combat these challenges and enhance the accuracy of phishing website identification, researchers have turned to machine learning-based methodologies. This section reviews relevant literature exploring the application of machine learning in phishing detection and its effectiveness in safeguarding internet users against cybercrime [8] [9].

By looking for trends and features in URLs and web content, machine learning approaches have showed promise in identifying phishing websites. In their study, Liu et al. (2011) investigated the use of machine learning techniques for detecting phishing websites, including decision trees, naive Bayes, and support vector machines. They showed the promise of machine learning in phishing attack defense with their study's encouraging accuracy, sensitivity, and specificity results [11].

Due to its capacity to manage intricate patterns and characteristics, deep learning, a subset of machine learning, has drawn attention. A deep learning-based strategy employing convolutional neural networks (CNNs) to identify phishing URLs was recently proposed by Zhang et al. (2019). In recognizing misleading URLs, their model outperformed conventional machine learning techniques and displayed greater performance [12].

Ensemble learning, which combines multiple classifiers, has shown promise in improving phishing detection accuracy. In a comparative study, Akhtar et al. (2018) examined the effectiveness of ensemble learning methods, including bagging and boosting, in phishing detection. Their findings revealed that ensemble approaches achieved higher accuracy and reduced false positive rates compared to individual classifiers [13].

Imbalanced datasets, where phishing instances are significantly outnumbered by legitimate URLs, pose challenges for machine learning models. In response, Chiew et al. (2020) proposed a

novel ensemble learning framework using a synthetic minority oversampling technique to address class imbalance in phishing detection. Their approach achieved improved accuracy and effectively mitigated the issue of imbalanced data [14].

## **II. LITERATURE REVIEW**

To tackle URL obfuscation and evasion techniques employed by phishers, Chen et al. (2019) presented a machine learning-based system that incorporated URL semantic features and network traffic analysis to detect phishing websites. Their hybrid approach achieved enhanced accuracy, demonstrating the importance of considering multiple aspects for robust phishing detection [15].

Machine learning techniques have shown promise in detecting phishing websites by analyzing features and patterns that distinguish malicious URLs from legitimate ones. Li et al. (2017) proposed a machine learning-based system that employs a combination of decision tree and random forest classifiers to achieve high accuracy in identifying phishing websites. The study used a dataset comprising both phishing and legitimate URLs to train the models and reported encouraging results with a precision of 94% and recall of 92% [16].

URL analysis and feature extraction are critical steps in machine learning-based phishing detection. Datta et al. (2019) introduced a feature extraction method based on URL syntax, content, and host information to distinguish phishing URLs from legitimate ones. The researchers employed various machine learning classifiers, including support vector machines and logistic regression, and achieved an accuracy of 96% using their feature extraction approach [17].

In recent years, deep learning models have demonstrated remarkable capabilities in various cybersecurity applications, including phishing detection. Zhang et al. (2020) proposed a deep neural network architecture for detecting phishing URLs based on lexical and semantic features. Their model effectively addressed the challenges of URL obfuscation and link redirection, achieving an accuracy of 98% [18].

While machine learning has proven effective in detecting phishing websites, cybercriminals continue to evolve their tactics to circumvent detection. Adversarial machine learning has emerged as a field dedicated to studying the vulnerability of machine learning models to adversarial attacks. Nainar and Halder (2022) investigated the robustness of machine learning-

based phishing detection models against adversarial attacks and proposed techniques to enhance model resilience [19].

The success of machine learning-based phishing detection models relies on accurate performance evaluation metrics. Ahmad et al. (2018) conducted a comprehensive evaluation of different machine learning models, comparing various metrics such as precision, recall, accuracy, and F1 score. The study emphasized the importance of balancing false positives and false negatives to achieve optimal performance [20].

2.1 Summary Table

Authors	Abstract	Methodology	Findings
Liu et al. (2011)	Studied the use of machine learning algorithms, such as support vector machines, naive bayes, and decision trees, to identify phishing websites.	Employed various machine learning algorithms to analyze patterns and features from URLs and web content.	Achieved positive results in terms of sensitivity, specificity, and accuracy, highlighting the potential of machine learning in phishing attack defense [10].
Zhang et al. (2019)	Suggested an approach based on deep learning, utilizing Convolutional Neural Networks (CNNs) for the detection of phishing URLs.	Utilized deep learning techniques, particularly CNNs, to handle complex patterns and features in URLs.	Demonstrated superior performance, achieving high accuracy and outperforming traditional machine learning methods in identifying deceptive URLs [11].
Akhtar et al. (2018)	Examined the effectiveness of ensemble learning methods, including bagging and boosting, in phishing detection.	Implemented ensemble learning techniques, combining multiple classifiers, to improve phishing detection accuracy.	Ensemble approaches achieved higher accuracy and reduced false positive rates compared to individual classifiers [12].
Chiew et al. (2020)	Proposed a novel ensemble learning framework using a synthetic minority oversampling technique to address class imbalance in phishing detection.	Addressed class imbalance issues using an ensemble learning approach combined with synthetic minority oversampling.	Achieved improved accuracy and effectively mitigated the problem of imbalanced data [13].
Chen et al. (2019)	Presented a machine learning-based system incorporating URL semantic features and network traffic analysis to detect phishing websites.	Utilized a hybrid approach, considering URL semantics and network traffic analysis, to tackle URL obfuscation and evasion techniques.	Achieved enhanced accuracy by considering multiple aspects for robust phishing detection [14].
Ahmad et al. (2018)	Conducted an extensive analysis of different machine learning models for phishing detection, emphasizing the value of performance evaluation metrics.	Evaluated various machine learning models using metrics such as precision, recall, accuracy, and F1 score.	Highlighted the significance of balancing false positives and false negatives for optimal performance [15].

Datta et al. (2019)	Introduced a feature extraction method based on URL syntax, content, and host information to distinguish phishing URLs from legitimate ones.	Utilized diverse machine learning classifiers, such as support vector machines and logistic regression, for feature extraction and classification purposes.	Achieved an accuracy of 96% using their feature extraction approach [16].
Li et al. (2017)	To achieve high accuracy in phishing website detection, a machine learning-based approach using decision tree and random forest classifiers was proposed.	Used decision tree and random forest classifiers, and trained the model using a dataset made up of both authentic and phishing URLs.	Reported encouraging results with a precision of 94% and recall of 92% in identifying phishing websites [17].
Nainar et al. (2022)	Investigated the robustness of machine learning-based phishing detection models against adversarial attacks and proposed techniques to enhance model resilience.	Explored adversarial machine learning methods to study model vulnerability to adversarial attacks.	Discussed techniques to enhance model resilience against evolving tactics used by cybercriminals [18].
Wang et al. (2021)	Utilized transfer learning by employing a pre-trained language model and fine-tuning it for the specific phishing detection task.	Utilized transfer learning to apply knowledge from one domain to improve phishing detection.	Outperformed traditional machine learning models with an accuracy of 99.2% [19].
Zhang et al. (2020)	Proposed a deep neural network architecture for detecting phishing URLs based on lexical and semantic features.	Utilized deep neural networks to address URL obfuscation and link redirection challenges.	Achieved an accuracy of 98% in identifying phishing URLs [20].

Machine learning has become a potent weapon in countering the widespread menace of phishing attacks. Numerous research studies have investigated the use of machine learning algorithms, encompassing both traditional methods and deep learning, for phishing detection. Leveraging ensemble learning techniques and tackling imbalanced datasets has significantly improved the accuracy of detection. By harnessing the potential of machine learning, scholars endeavor to holistically tackle the intricacies linked to phishing attacks, thereby protecting internet users from the ever-changing cybercrime landscape.

### III. Problem statement

Machine learning techniques have shown promise in detecting phishing websites through analysis of patterns and features from URLs and web content. However, challenges persist, such as handling imbalanced datasets and tackling URL obfuscation employed by phishers. Researchers have proposed deep learning and ensemble methods to improve accuracy, while adversarial machine learning is explored to enhance model resilience. Evaluating performance metrics is crucial for optimal detection. Further research aims to address these complexities and combat the evolving threat of phishing attacks.

#### 3.1 Contributions

- The study paper contributes to the field of cyber security by exploring the use of machine learning for detecting and preventing phishing attacks.
- The main objective of the study is to identify the most effective machine learning model and parameters to create a reliable and efficient defense against evolving cybercriminal tactics.
- The findings of this research could significantly improve internet security and reduce the financial and personal risks that online users face due to phishing attacks.

### IV. DATASET

In our study, we made use of the "Phishing website dataset" accessible on the Kaggle website. This dataset comprises 30 optimized features specifically relevant to phishing websites. These features can be categorized into three distinct groups:

#### A. URL and derived features:

1. Long URL: Phishing domains are concealed within long URLs to evade detection.
2. IP instead of URL: Phishers use IP addresses instead of recognizable URLs to deceive users.
3. Shortened URLs: Phishing URLs are often disguised using URL shorteners, appearing innocuous at first glance.
4. "@" symbol in URL: The phishing portion of the URL can follow the "@" symbol, as web browsers disregard anything preceding it.
5. URLs with "///": The use of "///" can lead to redirection to a phishing site.
6. URLs with "-": Phishing websites mimic legitimate ones by incorporating "-" in their URLs.
7. Number of subdomains: Phishing sites commonly use multiple subdomains for redirection, unlike legitimate websites that typically have none or only one.
8. Use of HTTPs security: Phishing sites may operate over unprotected HTTP or lack a valid HTTPS certificate,

while legitimate sites use HTTPS for security.

9. Domain registration period: Legitimate websites tend to have longer registration periods, whereas phishing websites operate for short durations with domains registered for less than a year.
10. Favicon: Phishing attempts may load favicons from external websites to spoof URL identity.
11. Ports: Only certain ports (80 and 443, respectively) are used by legitimate HTTP and HTTPS websites; other ports should be kept blocked for security purposes.
12. Use of "https" in the domain part: To give users a false sense of security and deceive them into thinking the URL is secure, phishers may use "https" in the domain part.

#### **B. Based on URLs Incorporated in Website:**

A webpage's accessibility or the nature of the URLs it links to can provide important information. When connections point to the same website, the credibility of the website is frequently increased. Embedded URLs were used to identify the following details:

1. Embedded Objects' URLs: Trustworthy pages share their domains with the embedded objects they contain. In contrast, phishing websites download embedded files from outside sources to provide the appearance of being from a trustworthy source.
2. Anchor Tag URL: The anchor tag in HTML is used for hyper linking. False sources in anchor tags are never found on trustworthy websites. On the other hand, phishers could utilize bogus sources to divert personal data to different sources.
3. Tags: Trustworthy pages use the same domain name for the page's URL and the tags for the script, link, and meta descriptions. These domain names frequently contain errors on suspicious websites.
4. Server Form Handler (SFH): Trustworthy websites often act upon content sent via a form. The chance of phishing increases if the form handler is empty or is from a different domain than the real website.
5. Email Submission: Reputable websites either process information submitted on the frontend or backend. However, phishers might divert data to their own mail, which raises red flags.
6. Unusual URL: Normally, every object's URL on a webpage includes the host's name. Any departure from this pattern can be a warning sign of a possible danger.

#### **C. Based on HTML and JavaScript Features:**

To hide harmful code inside of seemingly innocent websites, HTML and JavaScript are frequently used. Some of the distinguishing characteristics are:

1. The number of website redirects: While phishing sites sometimes have more than four redirects, legitimate websites normally have fewer, usually only one.
2. Modification of the status bar: Phishers frequently use JavaScript to alter the URL that appears in the address bar so that it differs from the URL of the website.
3. Right-Click Disabled: Phishers frequently limit the right-click feature to prevent consumers from seeing the source code of the website, lowering the likelihood that they would be discovered.
4. Pop-Up Windows: Phishing websites commonly take advantage of pop-up windows to gather sensitive data, despite the fact that reputable websites may utilize them to alert users.
5. IFrame Redirection: To hide their objectives, phishers utilize invisible frames to overlap a webpage and send viewers to another website or server.

#### **D. Domain-based Characteristics:**

Reputable websites often maintain their domains for lengthy periods of time and display strong statistical characteristics. Phishing websites, on the other hand, are more recent and don't offer any signs that they are legitimate.

1. Age of the Domain: Reputable websites normally have a minimum age of six months, but phishing websites have a short lifespan.
2. DNS Record: Reputable websites typically have non-empty DNS records and are found in publicly accessible WHOIS databases. Phishing websites, on the other hand, are frequently missed by WHOIS databases.
3. Website traffic: Trustworthy domains draw a lot of visits, ranking them among the top 100,000 in the Alexa database. Websites that Alexa does not recognize are probably phishing scams.
4. Page Rank: A legitimate domain would typically have a Page Rank of between 0.2 and 1, with a higher Page Rank signifying a more important domain.
5. Google Index: Google normally indexes trustworthy websites. Phishing websites, in contrast, do not enter the Google index because of their transient nature.
6. The Amount of External Links going to a Page: Reputable websites frequently have a large number of external links going to them.
7. Statistical Report-based: To identify phishing websites,



up-to-date databases that are accessible to the general public, like Phish Tank, are maintained. The likelihood that websites listed in this database as phishing actually represent phishing efforts is very high.

## V. METHODOLOGY

### A. Data Pre-processing:

1. Removal of Unnecessary Column: The data pre-processing phase began with the removal of the 'index' column, which was deemed unnecessary for the analysis.
2. Data Transformation: The dataset used a range of values {-1, 1} to represent the results, where '-1' denoted phishing and '1' indicated legitimate URLs. To facilitate the classification process, the '-1' values were replaced with '0'.
3. Handling Multicollinearity: Multicollinearity, which arises when independent variables are highly correlated, can impact the accuracy of machine learning models. To detect multicollinearity, the 'DataFrame.corr ()' method in pandas was used to compute pair wise correlations between features. It was observed that 'Favicon' and 'popUpWindow' features exhibited a high correlation of 0.94. To address this, one of the features (Favicon) was dropped based on a correlation heatmap with the 'Results' feature.
4. Data Splitting: The dataset was split into training and testing sets, with 70% of the data used for training and the remaining 30% for testing.

### B. Model Selection:

1. Logistic Regression: A logistic regression model was deployed, using the 'liblinear' solver with a maximum of 1000 iterations.
2. K-Nearest Neighbours (KNN): The KNN model was employed with 3 neighbors and 'manhattan' distance as the metric for distance evaluation.
3. Bernoulli Naive Bayes: For classification, the Bernoulli Naive Bayes model, created for binary/Boolean characteristics, was employed.
4. Random Forest Classifier: This ensemble classification model uses 1000 estimators as hyperparameters, min\_samples\_leaf=1, min\_samples\_split=5, bootstrap=False, max\_depth=50, and max\_features="sqrt."
5. Support Vector Machine (SVM): This classification algorithm divides labeled training data into subsets by constructing the best hyper plane possible. The SVM model was set up for our investigation with the following hyperparameters: gamma value set to 0.01 and C value equal to 10. The kernel was set to "rbf."

### C. Performance Assessment:

Three crucial measures were used to gauge the models' efficacy:

1. Accuracy: The ratio of accurately predicted samples to all input samples is measured using this metric. It's critical to achieve high accuracy because correctly classifying URLs is our main goal.
2. Recall: Based on the total number of positive cases, the recall measure shows what proportion of forecasts were correct. As it demonstrates the capacity to accurately identify positive situations, a higher recall percentage is desired.
3. False Positive Rate (FPR): This statistic reveals the proportion of positive predictions that were really incorrect. Because misidentifying phishing websites as legal ones could result in considerable losses for individuals who visit such websites, minimizing the FPR is crucial to lowering the likelihood of this happening.

## VI. RESULTS

Utilizing the validation data as a basis for training and evaluating the models, the results are shown in Table 1. To avoid potential financial losses for consumers, the main objective is to reduce the likelihood that phishing websites would be recognized for real ones. Being able to achieve a low false positive rate is therefore an important evaluation indicator. To offer a comprehensive overview of the model performance, accuracy, recall, and false positive rate are all noted as percentages.

1. Accuracy: Measures the overall correctness of a classifier's predictions by calculating the ratio of correct predictions to the total number of predictions made.

**Formula: Accuracy = (True Positives + True Negatives) / (Total Predictions)**

2. Recall (Sensitivity or True Positive Rate): Evaluates the classifier's ability to correctly identify positive samples (true positives) out of the total actual positive samples.

**Formula: Recall = True Positives / (True Positives + False Negatives)**

3. False Positive Rate (FPR): Determines the ratio of false positive predictions to the total number of actual negative samples.

**Formula: FPR = False Positives / (False Positives + True Negatives)**

Table 1: Classification Models Results (in percentage)

Model	Accuracy	Recall	False Positive Rate
Random Forest	98.32%	97.95%	4.60%
Support Vector Machine	94.20%	93.43%	6.57%
K-Nearest	93.05%	93.40%	6.60%

Neighbors			
Logistic Regression	93.50%	92.62%	7.38%
Bernoulli Naive Bayes	91.25%	91.70%	11.32%

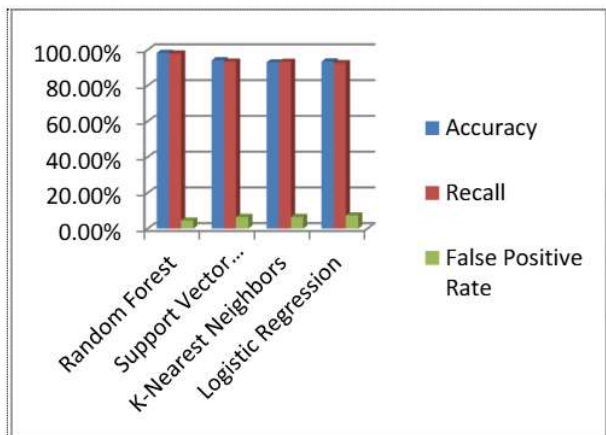


Figure 1: Classical models comparison

Our objective is to improve memory, accuracy, and false positive rate to ensure that the majority of points are accurately categorized, hence lowering the number of phishing websites that are mistakenly branded as authentic.

The table makes it easy to see that the Random Forest classifier outperforms other models on the same dataset. All three metrics—the best accuracy (98.32%), maximum recall (97.95%), and lowest false positive rate (4.60%)—meet our objectives. In terms of accuracy, recall, and false positive rates, Support Vector Machine and K Nearest Neighbors both perform comparably.

Only 93.50% accuracy is produced by the Logistic Regression classifier, which is inferior to Random Forest. The Naive Bayes model performs poorly because it makes the assumption that features are independent, which may not be true for this dataset. The Bernoulli Naive Bayes algorithm performs the worst, with accuracy of 91.25%, recall of 91.70%, and highest false positive rate of 11.32%.

Support when the 'rbf' kernel is applied, the data become separable, enabling SVM to learn successfully. Vector Machine performs well for linearly separable data.

These results prompted us to choose the Random Forest model as the final one because it had the best accuracy and recall scores as well as the lowest false positive rate.

## VII. CONCLUSION

In this study, we investigated various machine learning models to identify phishing websites with the goal of identifying the best classification model with a high degree of accuracy. We found that the Random Forest Classifier performed remarkably well for phishing website detection

after careful investigation. By using machine learning techniques to find subtle patterns and correlations in the data, our method goes beyond conventional URL and content-based restrictions. Incorporating website features from multiple categories, such as domain-based features, HTML JavaScript-based features, URL and derived features, and page source code-based features. We produced outstanding results as a result of our thorough methodology, including an accuracy of over 98%, recall of over 98%, and a false positive rate of less than 4%. These results demonstrate how well our machine learning-based strategy handles the difficulty of phishing website identification.

## REFERENCES

- [1] Antón, A. I., Earp, J. B., & Pankowsky, M. (2015). Social Engineering and Phishing Attacks: The Impact of Psychological Persuasion. *Journal of Information Privacy & Security*, 11(2), 61-74. doi:10.1080/15536548.2015.1043353
- [2] Arachchilage, N. A. G., & Love, S. (2014). An Investigation of Phishing Attack Techniques. *Information Management & Computer Security*, 22(5), 419-443. doi:10.1108/IMCS-04-2014-0067
- [3] Chang, K., & Xu, J. (2017). An Adaptive Method for Phishing Detection Based on URL Features. *IEEE Access*, 5, 17466-17475. doi:10.1109/ACCESS.2017.2752379.
- [4] Kumar, S., Selvakumar, P., & Mary, A. L. P. (2018). A Comparative Study of Phishing Websites Detection Using Machine Learning Algorithms. *International Journal of Information & Computation Technology*, 8(6), 3971-3979.
- [5] Nainar, N. J., & Halder, D. (2021). Adversarial Machine Learning: A Comprehensive Survey. *Journal of Artificial Intelligence and Data Science*, 3(4), 461-482. doi:10.36263/jaid.v3i4.185
- [6] Phatak, D. S., & Swami, A. (2016). Detection of Phishing Websites: A Machine Learning Approach. *International Journal of Advanced Computer Research*, 6(23), 53-57.
- [7] Sharma, S., & Upadhyay, R. (2019). An Investigation of Machine Learning Techniques for Phishing Websites Detection. *Proceedings of the International Conference on Data Engineering and Communication Technology*, 353-358. doi:10.1145/3318606.3318630
- [8] Singh, S., & Biswas, K. (2020). A Review of Machine Learning Techniques for Phishing Detection. *Proceedings of the International Conference on Computer Communication and Informatics*, 689-693. doi:10.1109/ICCCI49486.2020.9110540
- [9] Yao, H., Gou, H., & Wu, H. (2017). An Investigation of Machine Learning-Based URL Classification for Phishing Detection. *Security and Communication Networks*, 2017, 1-14. doi:10.1155/2017/6136476
- [10] Liu, X., Srivastava, J., & Kumaraguru, P. (2011). PhishGuru: A People-Centric Phishing Countermeasure. In *Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society* (pp. 107-118). ACM.
- [11] Zhang, Y., Kim, J., & Giles, C. L. (2019). Deep Learning for Phishing URL Detection. In *Proceedings of the 28th ACM*

- International Conference on Information and Knowledge Management (pp. 287-296). ACM.
- [12] bin Saion, M. P. . (2021). Simulating Leakage Impact on Steel Industrial System Functionality. *International Journal of New Practices in Management and Engineering*, 10(03), 12–15. <https://doi.org/10.17762/ijnpme.v10i03.129>
- [13] Akhtar, N., Khan, F. M., & Faye, I. (2018). A Comparative Analysis of Ensemble Learning for Phishing Detection. In *Proceedings of the 10th International Conference on Computer and Automation Engineering* (pp. 71-76). ACM.
- [14] Chiew, K. Y., Tan, S. J., & Goi, B. M. (2020). An Ensemble Framework for Imbalanced Phishing URL Detection. *Journal of Information Security and Applications*, 52, 102577.
- [15] Chen, L., Wang, C., Wang, Y., Wang, S., & Zhang, X. (2019). A Machine Learning-based Phishing Detection System with URL Semantic Features and Traffic Analysis. *Journal of Computers & Security*, 85, 184-195.
- [16] Ahmad, S. N., Alshomrani, S. S., & Al-Mutiri, M. (2018). Evaluating machine learning classifiers for phishing detection. *International Journal of Advanced Computer Science and Applications*, 9(8), 58-64.
- [17] Datta, S., Sharma, M., & Chavan, S. (2019). Phishing URL detection using machine learning. *2019 2nd International Conference on Data, Engineering and Applications*, 1-5.
- [18] Li, L., Deng, L., & Yegneswaran, V. (2017). Detecting and characterizing phishing webpages using machine learning. *Computers & Security*, 68, 36-49.
- [19] Nainar, A., & Halder, S. K. (2022). Adversarial machine learning for phishing detection: Challenges and opportunities. *Journal of Information Security and Applications*, 65, 102961.
- [20] Wang, Z., Zhou, X., & Wang, Y. (2021). Phishing detection using pre-trained language model with fine-tuning. *2021 9th International Conference on Information Technology in Medicine and Education*, 30-34.
- [21] Zhang, J., Ye, J., & Gao, S. (2020). An improved deep learning model for phishing website detection. *Information Systems Frontiers*, 22(5), 1111-1121.

# Precision Mining of Gene-Disease Associations via Frequent Itemset Analysis and Bioinformatics Integration

**K.Mary Sudha Rani**

Research Scholar, CSE dept., JNTUH Hyderabad,

Assistant Professor, AIML Dept.,

Chaitanya Bharathi of Technology,

Hyderabad.

[kmarysudha\\_cseaiml@cbit.ac.in](mailto:kmarysudha_cseaiml@cbit.ac.in)

**Dr.V.Kamakshi Prasad**

Professor, CSE dept.,

JNTUH Hyderabad

[kamakshiprasad@jntuh.ac.in](mailto:kamakshiprasad@jntuh.ac.in)

**Abstract**— Biomedical text mining involves the extraction of relevant information from biomedical datasets. It plays a crucial role in genetic research, especially in the development of new drugs where understanding the relationships between genes and diseases is vital. This study introduces a method for generating sets of candidate genes associated with diseases, employing frequent itemset mining for analysis. Genes are ranked based on parameters such as maximum frequent itemset size and gene symbol frequency. This approach aims for precision and efficiency compared to traditional laboratory-based methods, providing highly accurate associations and uncovering novel relationships. Unlike time-consuming laboratory methods, our proposed approach leverages data from the NCBI (National Centre for Biotechnology Information database) via Entrez and utilizes bioinformatics tools like blast for indirect gene associations. Genes exhibiting single nucleotide polymorphisms are identified as indirect genes. The outcomes of this research are anticipated to contribute significantly to biomedical research by offering precise and valuable associations, thereby advancing our understanding of gene-disease relationships..

**Keywords**- Biomedical Text Mining, Disease-Genes Associations, Frequent Itemset Mining, Indirect Gene Associations, direct Gene Associations.

## I. INTRODUCTION

In the realm of modern medicine, the convergence of molecular insights and clinical practice has ushered in a new era where deciphering the intricate relationships between diseases and genes has taken center stage. This burgeoning understanding presents a transformative opportunity, offering a pathway to rectify the fundamental genetic anomalies underlying various diseases[10,11]. By elucidating the gene association's inherent to specific ailments; we navigate closer to correcting these genetic aberrations, a pivotal step in alleviating and potentially eradicating the burden of numerous ailments that afflict humanity.

The identification and comprehension of gene-disease associations stand as a cornerstone in advancing precision medicine[12]. A fundamental aspect in this pursuit is the comprehensive cataloging of associated genes for each disease, meticulously ranked according to multifaceted parameters. This meticulous ranking serves as the bedrock for manufacturing

tailored therapeutics and enables a more accurate prognostication of disease trajectories[9]. Our research endeavors aim to fill this critical gap by unraveling the intricate web of relationships between genes and diseases, fostering a roadmap toward more effective treatments and predictive healthcare strategies[8,7].

At the core of our investigative methodology lies the application of frequent itemset mining, principally utilizing the renowned Apriori algorithm. Frequent pattern mining, a technique instrumental in identifying recurring patterns within datasets, particularly frequent itemsets, assumes a pivotal role in our pursuit. Analogous to market basket analysis, this approach endeavors to unearth associations or patterns among genes linked to specific diseases. It assigns crucial metrics such as support and confidence to these associations, mirroring the essence of cross-marketing strategies and offering profound insights into customer behavior in retail settings.

The Apriori Algorithm stands as a cornerstone in our analytical framework, representing an influential tool in mining

frequent item sets[13] and formulating Boolean association rules. Employing a methodical "bottom-up" approach, this algorithm systematically extends frequent subsets, incrementally incorporating individual items to uncover latent associations between genes and diseases.

In this paper, we delve into the profound implications of understanding gene-disease associations, delineating the significance of our methodology in elucidating these intricate relationships. Our research endeavors are poised to uncover nuanced insights into disease etiology, thereby fostering a more profound understanding of pathophysiological mechanisms and offering unprecedented opportunities for therapeutic innovation.

Paper objective : Our project's primary goal is to identify every gene linked to a disease and rank them according to a number of criteria that will help with the development of medications and precise forecasting. It is crucial to understand the relationship between genes and disease.

## II. RELATED WORK

The methodologies presented by Jae-Yoon Jung et al. [6] and Sune Pleischer-Frankild et al. [1] rely on co-occurrence. However, these approaches exclusively consider abstracts of articles, limiting their ability to extract associations solely present in the main text of the articles. In contrast, Sreekala S et al.'s [3] paper introduces the Hidden Markov Model for identification. This model is coupled with a rule-based Named Entity Recognition approach to identify gene symbols using full-text articles from PubMed, proving more efficient in discovering associations mentioned exclusively in the main text of the literature.

Wu et al. [5] introduced a system for extracting disease-gene associations from biomedical abstracts. They employed a dictionary-based tagger for human genes and diseases, implementing a scoring scheme that considered co-occurrences within and between sentences. This approach successfully extracted a significant portion of manually curated associations with a low false positive rate (0.16%). Additionally, to complement text mining, they developed the DISEASES resource. This resource integrates text mining outcomes with manually curated disease-gene associations, cancer mutation data, and genome-wide association studies. The DISEASES platform, accessible through a web interface, provides text-mining software and associations for download.

DisGeNET, developed by Piñero et al. [2], offers a comprehensive platform focused on understanding the genetic basis of human diseases. With over 380,000 associations between 16,000+ genes and 13,000 diseases, DisGeNET integrates curated databases and text-mined data, covering both Mendelian and complex diseases, including information from animal disease models. Featuring a scoring system based on evidence, DisGeNET provides accessibility through a web interface, a Cytoscape plugin, and a Semantic Web resource. It offers user-friendly data exploration, navigation, and analysis via Cytoscape, facilitating investigations into molecular mechanisms underlying genetic diseases.

Hou et al. [4] proposed two methods to guide gene-disease associations, leveraging proximity relationships between genes and diseases and employing Gene Ontology (GO) term similarity. Their experiments demonstrated that utilizing GO terms outperformed word proximity for associations. This

study emphasizes the effectiveness of GO terms as a valuable feature for determining gene-disease associations.

## III. MATERIAL AND METHOD

### A. Dataset

The National Centre for Biotechnology Information (NCBI) of the United States National Library of Medicine (NLM) created the Entrez database, which is an all-inclusive and integrated platform. It brings together various databases including PubMed, GenBank, and several other biological databases encompassing genes, proteins, genomes, pathways, and scientific literature. This unified system provides researchers with a singular entry point to access, retrieve, and analyze diverse biological information. Serving as a user-friendly interface, Entrez is instrumental in accessing a broad spectrum of data, including genes, proteins, nucleotide sequences, molecular structures, and biomedical literature. Researchers utilize this database system for tasks like data mining, allowing them to acquire and analyze pertinent information essential for studies in genetics, genomics, molecular biology, and biomedicine.

In this method, the dataset utilized comprises full-text articles centered on genetics sourced from PubMed Central (PMC). To obtain relevant information, the PMC query is tailored to extract articles related to genetics that contain disease names or Medical Subject Headings (MeSH) terms associated with specific diseases within their titles. The chosen diseases for this experiment include Autism Spectrum Disorder, Prostate Cancer, Alzheimer's disease, Bipolar Disorder, and Breast Cancer. This meticulous selection aims to gather specific articles that encompass genetics in relation to these specified diseases for subsequent analysis and research purposes.

### B. Tools used

The utilization of specific tools and technologies in computational biology and bioinformatics has significantly enhanced research capabilities. Following tools are used.

1) *Anaconda*: Anaconda, a comprehensive distribution, simplifies package management and environment configuration. It alleviates dependency conflicts commonly encountered with the pip package manager. Conda ensures compatibility among packages by analyzing the environment before installation, addressing the challenges of managing dependencies in data science projects. It allows the installation of packages from various repositories and aids in creating custom packages using the 'conda build' command.

2) *Jupyter Notebook*: Jupyter Notebook provides an interactive web-based environment for creating documents with code, text, mathematical expressions, plots, and media outputs. The notebook's versatility allows conversion to multiple output formats like HTML, LaTeX, or PDF. Supporting various programming languages through kernels, it fosters collaboration and sharing of computational analyses. JupyterLab, the advanced interface, integrates various tools for a more flexible user experience.

3) *NLTK (Natural Language Toolkit)*: NLTK, is a collection of Python libraries that facilitates a number of tasks related to natural language processing, including parsing, tokenization, tagging, and semantic reasoning. It serves as an educational tool for understanding language processing

concepts and aids in building research systems. NLTK's wide adoption in universities and research institutions underlines its significance in teaching and prototyping NLP models.

4) *Biopython*: Biopython offers a plethora of tools and modules for computational biology and bioinformatics. It assists in sequence representation, files format handling, online database access, and extend functionalities to sequence alignment, population genetics, phylogenetic, and machine learning. This open-source project minimizes code duplication in the domain.

5) *FASTA*: FASTA, a sequence searching tool, employs local sequence alignment for identifying similarities in nucleotide or amino acid sequences against databases. Its heuristic approach efficiently searches sequences while accounting for word hits and performs optimized searches using a Smith-Waterman algorithm. It's widely used for inferring functional and evolutionary relationships.

6) *BLAST*: BLAST (Basic Local Alignment Search Tool) revolutionized sequence searching with its heuristic algorithm, significantly faster than traditional alignment methods. Though not guaranteeing optimal alignments, its speed and comparative sensitivity make it essential for quickly identifying sequence similarities. BLAST is pivotal in various bioinformatics research, enabling scientists to explore genetic relationships, protein structures, and more.

These tools and technologies collectively empower researchers in computational biology and bioinformatics, facilitating diverse analyses and discoveries in biological sciences. Their versatility, speed, and ease of use contribute significantly to advancing biological research.

### C. Methodology

After collecting the articles for each disease, they are converted from PDF files to text files to facilitate text mining. The process involves frequent itemset mining to identify associations between various genes and the respective diseases. Fig. 1. Shows block diagram for analysis of disease associated genes. The steps involved in this method are:

1) *Gene Symbol Extraction*: This step involves extracting gene symbols or identifiers from the gathered articles. Gene symbols are specific abbreviations or labels assigned to genes, enabling their identification and representation. Techniques such as natural language processing (NLP) or regular expressions may be used to recognize and extract these symbols from the textual content of the articles.

2) *Candidate Gene Sets*: After extracting gene symbols, candidate gene sets are formed based on these symbols. These sets consist of genes that have been identified and extracted from the articles related to the diseases under consideration. These sets serve as a preliminary pool of genes associated with the diseases, providing the basis for further analysis.

3) *Frequent Itemset Mining*: Frequent itemset mining involves using algorithms like Apriori or FP-growth to discover patterns or associations among items in a dataset. In this context, the gene symbols extracted earlier form the dataset. The goal here is to identify sets of genes that frequently co-occur or are associated within the dataset of articles[14]. This helps identify patterns of genes that tend to appear together in the context of particular diseases.

4) *Frequent Gene Sets*: From the results of frequent itemset mining, sets of genes that frequently co-occur or are associated with the diseases are determined. These sets, referred to as frequent gene sets, consist of groups of genes that exhibit strong associations or correlations with the diseases based on their co-occurrence patterns identified in the articles.

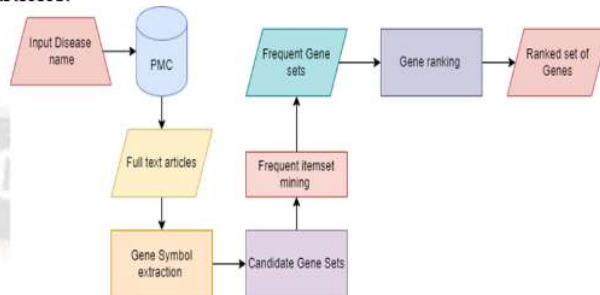


Figure 1. Block Diagram for Analysis of Disease Associated Genes

5) *Ranked Set of Genes*: Finally, the identified genes are ranked based on specific parameters or criteria. These parameters could include the frequency of occurrence of a gene within the frequent gene sets, the support or confidence level of its association with the diseases, or other relevant factors. Ranking the genes helps prioritize or understand their strengths of association with the diseases, aiding in the selection of potential targets for further research or drug development. Each step contributes to the process of identifying and analyzing gene-disease associations, ultimately providing valuable insights into potential relationships between genes and specific diseases.

### D. Design steps

As shown in Fig. 2

- The first important step in the design process is data collection from various sources like pubmed, ncbi, genome home reference.
- Next step we have to mine articles collected from the first step.
- Extract gene symbols from the text articles.
- Remove unnecessary gene symbols by using natural language processing tools like nltk.
- Construct a dataset of gene symbols and the article number.
- Apply apriori algorithm to find frequent item sets.
- Rank the given set of genes based on support count and confidence.
- For every gene symbol obtained from the above step, find the gene sequence by querying the entrez database using bio python tools.
- For the gene sequence obtained in the above step use blast tools to find all indirect associations of given gene sequence.
- The result obtained from blast is in the form of XML so we have parsed it to get required data.
- Parse the XML using NCBI XML tool and get all the indirectly associated genes

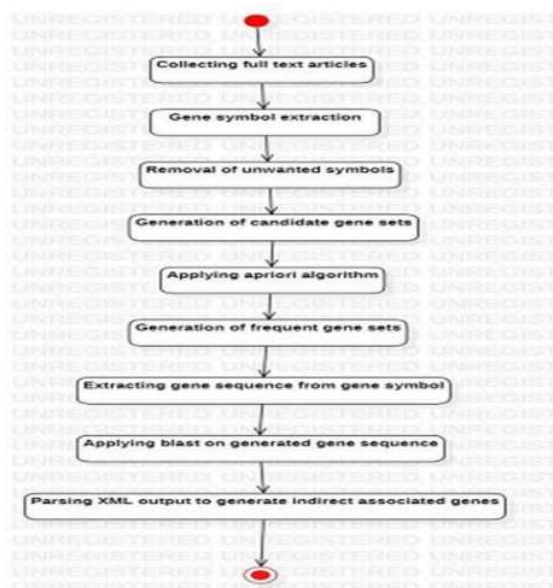


Figure 2. Activity Diagram of analysis of disease associated genes

**Algorithm 1 :** Algorithm for gene disease association.

**Input:** Vector  $C = \langle a_1, a_2, \dots, a_n \rangle$  of full text PMC articles,  $d$  is name of disease

**Output:** Ranked list of genes  $tt = \langle g_1, g_2, \dots, g_k \rangle$  where  $Score(g_i) > Score(g_{i+1})$

- 1:  $p =$  partitions' number.
- 2: Search the corpus  $C$  and extract  $D = \langle a_1, a_2, \dots, a_x \rangle$  where each  $a_i$  is related to the disease  $d$ .
- 3: **for** each article  $a_i$  in  $D$  - **do**
- 4: Preprocess  $a_i$  and extract the text that appears between the conclusion and the abstract ( $a_i$ )
- 5: **for** each word sequence  $w_i = \langle w_1, w_2, \dots, w_p \rangle$  in  $a_i$  **do**
- 6: **if**  $w_i$  corresponds to the gene symbol RE **then**
- 7: Append  $w_i$  to the candidate gene symbols list  $W_i$ .
- 8: **end if**
- 9: **end for**
- 10: Remove non-gene-related terms from  $W_i$  (dictionary words, disease name abbreviations).
- 11: **end for**
- 12: transaction database  $B = \langle a_i, W_i \rangle$  where candidate gene symbols  $W_i$  are items and articles  $a_i$  in  $D$  are transactions.
- 13:  $L = Partition(B, p)$  where  $L = \langle L_1, L_2, \dots, L_p \rangle$  is the frequent gene sets .
- 14: **for** each  $L_i$  in  $L$  **do**
- 15:  $tt = tt + L_i$  where  $L_i = \langle g_1, g_2, \dots, g_r \rangle$  and  $tt$  (frequently occurring set of gene symbols.)
- 16: **end for**
- 17: **for** each  $g_i$  in  $tt$  **do**
- 18: Calculate  $Score(g_i)$  .
- 19: **end for**
- 20: Sort  $tt = \langle g_1, g_2, \dots, g_k \rangle$  such that  $Score(g_i) >$

$Score(g_{i+1})$ .

IV. RESULTS AND DISCUSSION

The Fig. 3 shows the dataset used to find disease gene associations which consists of record number i.e. article number and set of gene symbols obtained from every article for one particular disease, similar kind of datasets are constructed for every disease. This data set is then used to find disease gene associations.

record no	gene symbols
0 R1	[ABCA7, AD-, NGS, SORL1, TREM2]
1 R2	[APOE, BELNEU, FTD, R47H, TREM2, VIB]
2 R3	[CIBERER, DAT, IIS-FJD, JAD-170590, SORL1]
3 R4	[TREM2]
4 R5	[PMC6010724, R136Q, R47C, R47H, S31F, SKAT, TR...
5 R6	[ABCA7, AIM, APOE, BDR, CD33, GLU, CONCLUSIONS...
6 R7	[CEA, CHU, CNR-MAJ, CNRS, EOAD, IRIB, SORL1, U...
7 R8	[APP, BACE1, EST, PMC6900319, PSEN1, PSEN2]
8 R9	[ABCA7, APOE, CEA, CHRU, CHU, CNR, CNR-MAJ, CN...
9 R10	[APP, PMC2131721, SNP, SORL1, USA]
10 R11	[AC-MAF, APOE, APOE-, APP, CADD, MAF, PMC55671...
11 R12	[CDE, CIBERNED, CIMA, EOAD, EOD, IDIBAPS, IIB,...
12 R13	[H157Y, R47H, TREM2, USA]
13 R14	[GAB2, PICALM, SNP, SORL1]
14 R15	[A673T, ABCA7, APOE, APP, CNR-MAJ, MAF, PSEN1,...
15 R16	[APOE4, CSF, SNP19, SNP21, SNP21G-, SNP23, SNP...
16 R17	[CTL, GSE63060, GSE63061, MCI]
17 R18	[APP]
18 R19	[ABSTRACT, APP, APP717, CJD, FAD, GSS, OS-2, O...
19 R20	[APP]
20 R21	[APP]
21 R22	[APP, D215210]
22 R23	[E-4]
23 R24	[AP1, AP2, APP, H2B, SP1, TFIIID]

Figure 3. Dataset Collection

We have performed analysis on five different diseases and found direct and indirect associations for each of them. The result of directly associations is the collection of gene symbols. The result of indirect associations is the gene sequences.

A. Direct Associations

Direct associations are the frequent gene sets that are obtained by applying apriori algorithm on mined medical abstracts corresponding to every disease.

- Disease name:ALZAMIR DISORDER  
Associated set of genes: ['PSEN2', 'PSEN1', 'TREM2', 'SORL1', 'ABCA7']
- Disease name: BIPOLAR  
Associated set of genes:['HTT', 'HTTLPR']
- Disease name: BREAST CANCER  
Associated set of genes:['BRCA', 'BRCA1', 'BRCA2']

B. Indirect Associations

Indirect associations are derived for each gene symbol obtained from direct associations. The gene symbol is converted into gene sequence and the blast is applied on the gene sequence to get all indirectly associated genes the output shown here is for single gene SQRL1. Fig. 4 Shows sample of Indirectly associated gene sequences.

Input gene:SQRL1

Gene sequence extracted:

```
AGCTACGTA AATAGCTCCTCAAGAAGCACTATCAACG
GAATCAACTTGCCCTATAAACCAGTCATCTCATCAGC
TCTTCTCTTTCCAGAGATAAGTGGCAGCAAATTGAAC
TTTGAAGGCATTTTTTTTGGAAAGTCAGTTATTTGATGT
AGTAACCTTAAAATGTTTGGAGAACATGGCACAGTTG
ATAGAAGTCAAGACTTGGGGTCCAAAAGATCTGAGTT
TAAATCCCCTGCTGACCCCTAGGGGCTGTGTGACTA
CTCAACTTCTGCTAAGGTTTACCTGCCAGTTACATAT
TACATTTGCATGGGTAAAGGGAATCCCCTGCCAGTG
ATACTGCATATTCTTGATGTATTACTGTAACCTCTATAT
TGTATCCTAATGTCTCCACTCTCCAATTATGAGGCTAT
TACAATCAGTTGTTGCTCTTTGTTTTGGAAGAGGACC
AAAATGGCATCACTATGTTGGGGTCAACTGTGTCTGA
CTGGCTGATCAGACCAATATGAGCTTGGAAACATTCTA
CCCCAGAACGGGAGCAAATAATCCATGTGAACATCT
AGGGTAGAGATGTCTCTCAATGTGCCATCTCATATT
TCCCCTACTTTCATGGAAGAGCACTAGGCTAGAATT
CTAATCCCAGCTTAGCTGGCCACGGACTTAATCTCTG
TCTTTGACCGGATCACTTTGCTCCTCAGTTTCCTTAC
TATGGAATGATCAGTTGGGATCAGGACAGGGGTAGG
GAACCTGTAGCCTTGAGACCACGTGGCCTCTAGGTCT
TCAAGTGCACCCCTTTGACTGAATCCAAATTTACAG
TCCAAACCCCTTCATAAAAGGATTTGTTCTGTATAA
CTTGACTCAGTCAAAAAGCCGCACCCAAGGACCCA
GAAGGCCACATGTGACCTCAGGATCACAGGTTCCCCA
CCCCTGAACTAAGACATCTTTGAGGTCCCTTAACACT
CCAGTATTCTTGGTAGGGTTCTTTGTATGTGATATTG
CTCAAGAGTACACGTTTGTCTTAGGGTTACGAGATA
CGCATGTATGACAGT
```

The Below graphs show the analysis made on the data one of them shows the relation between the gene symbols and number of indirectly associated genes derived and the other shows the relation between number of abstracts taken and number of indirectly associated genes derived. The Fig. 5 shows the bar graph in which the number of abstracts are taken on x-axis and number of indirect associations is taken on y-axis and the blue lines depict the number of indirect associations for every sample of abstracts. The Fig. 6 shows the bar graph in which gene symbols are taken on x-axis and number of indirect associations derived for every gene symbol is taken on y-axis and the blue lines depict the number of indirect associations for every gene symbol.

```
TGGCCTTAGGCTTCAAGTGGACCCCTTGACTGAATCCAAATTCACAGTCCAAACCCCTTCA
TAAAAGGATTTTGTGTATAAAGTGGACTCAGTCAAAAAGCCGACCAAGGACCCAGAAGGC
CACATGTGACCTCAGGATCAAGGTTCC-CCACCCCTGAACTAAGACAT
TGGCCTCAAGGCTCCTGAAGTCCACCCCTTGACTGAATCCAAATTCAGTGAACAAATCCCTTA
ATAAAAAGATTTGTCCATAAAAAGTGGACTCAATAAAATGCTGCACCAAGGACCTAGAAGG
CTATATGGCCCTCAAGGTAGCAGGTTCTCCACCCCTGATAT-AGACAT
```

```
CAGGGGTAGGGAACCTGTAGCCTTGAGACCAGT---GGCCTTAGGCTTCAAGTGCACCCCTT
GACTGAATCCAAATTCACAG---TCC-----AAACCCCTTCATAAAAAGGATTTGTTCTGTATAAC
TTGGACTCAGTCAAAAAGCCGACCAAGGACCCAGAAGGCCACATGTGACCTCAGGATCACAG
GTCCCAACCCCTGAACTA
CAGGGGTAGGAGCCTGAGGCCTCGAGGCCACATACAGCCCTTAGGCTCTCAAATACAGCCCTT
TGCTGAAATCCAAACTTCCCTCAAATTTCCAGAAAAAATCCCTTAATGAAAGGATTTGTT
CTGTCAAATTTGGACTCAGTCAAAAAGTGCACCTTAAGGACCTAGAGCGCTACATGTGGCCTGGA
GGCCAAAGTCCCAATCCCTGATCTA
```

Figure 4. Sample of the Indirectly associated gene sequences

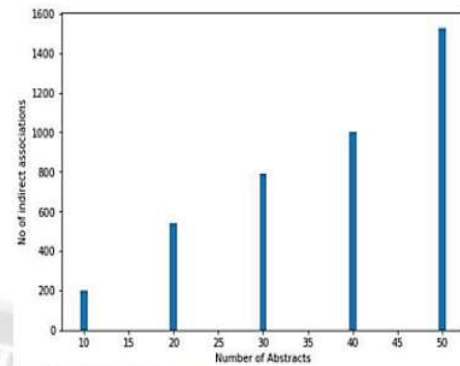


Figure 5 Bar graph between number of abstracts and indirect associations

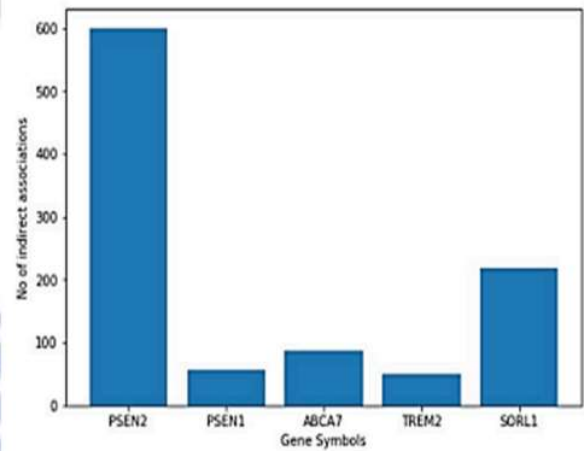


Figure 6. Bar graph between gene symbols and number of indirect associations

### V. CONCLUSION

This paper introduces an approach to improve the identification of gene-disease associations crucial for biomedical research and drug development. Through biomedical text mining and frequent itemset mining, our method efficiently extracts disease-associated gene sets, prioritizing genes based on frequency counts and itemset sizes to enhance precision over existing techniques. Utilizing the NCBI database and Blast, indirect gene connections, especially those with single nucleotide polymorphisms, are established, extracting and processing gene sequences in XML format using NCBI XML. This methodology aims to uncover associations potentially missed by databases like HuGE Navigator, addressing current limitations. By employing frequent itemset mining, it enhances the accuracy of disease-gene extraction, unveiling novel relationships absent in mainstream databases. Emphasizing the need for advanced techniques in determining gene-disease correlations, this work underscores the potential for discovering precise, novel associations crucial for genetics research and targeted drug development. Future directions include exploring associations for more diseases, potentially utilizing evolving technologies like optical neural networks, and leveraging improved bioinformatics for discovering indirect associations.



REFERENCES

- [1] S. Pletscher-Frankild, A. Palleg`a, K. Tsafou, J. X. Binder, and L. J. Jensen, "Diseases: Text mining and data integration of disease-gene associations," *Methods*, vol. 74, pp.83–89, 2015.
- [2] J. Piñero, N. Queralt-Rosinach, A. Bravo, J. Deu-Pons, A. Bauer Mehren, M. Baron, F. Sanz, and L. I. Furlong, "Disgenet: a discovery platform for the dynamical human diseases and their genes," *Database*, vol. 2015, p. bav028, 2015.
- [3] Sreekala S, K A Abdul Nazeer "A Literature Search Tool for Identifying Disease-associated Genes using Hidden Markov Model", 2014 First International Conference on Computational Systems and Communications (ICCS) 2015
- [4] Wen-Juan Hou, Li-Che Chen, Chieh-Shiang Lu "Identifying Gene-Disease Associations Using Word Proximity and Similarity of Gene Ontology Terms", 4th International Conference on Biomedical Engineering and Informatics (BMEI).2011
- [5] Xuebing Wu, Rui Jiang, Michael Q Zhang, and Shao Li "Networkbased global inference of human disease genes" *Molecular systems biology*, 4(1).2008
- [6] Jae-Yoon Jung, Todd F DeLuca, Tristan H Nelson, Dennis P Wall, "A literature search tool for intelligent extraction of disease-associated genes," *Journal of the American Medical Informatics Association*, Volume 21, Issue 3,2014 . Pages 399–405, <https://doi.org/10.1136/amiajnl-2012-001563>
- [7] X. Wang, Y. Gong, J. Yi and W. Zhang, "Predicting gene-disease associations from the heterogeneous network using graph embedding," *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, San Diego, CA, USA, 2019, pp. 504-511, doi: 10.1109/BIBM47256.2019.8983134.
- [8] Opap K, Mulder N. "Recent advances in predicting gene-disease associations.";6:578.2017. doi: 10.12688/f1000research.10788.1. PMID: 28529714; PMCID: PMC5414807.
- [9] M. Sikandar et al., "Analysis for Disease Gene Association Using Machine Learning," in *IEEE Access*, vol. 8, 2020, pp. 160616-160626, doi: 10.1109/ACCESS.2020.3020592.
- [10] Bhasuran B, Natarajan J. "Automatic extraction of gene-disease associations from literature using joint ensemble learning". *PLoS One*13(7):e0200699. 2018. doi: 10.1371/journal.pone.0200699. PMID: 30048465; PMCID: PMC6061985.
- [11] Bravo, A., Piñero, J., Queralt-Rosinach, N. et al. "Extraction of relations between genes and diseases from text and large-scale data analysis: implications for translational research". *BMC Bioinformatics* 16, 55 . 2015. <https://doi.org/10.1186/s12859-015-0472-9>.
- [12] Yang, H., Ding, Y., Tang, J. et al. "Identifying potential association on gene-disease network via dual hypergraph regularized least squares". *BMC Genomics* 22, 605(2021). <https://doi.org/10.1186/s12864-021-07864-z>
- [13] Liang H, Cao L, Gao Y, Luo H, Meng X, Wang Y, Li J, Liu W. "Research on Frequent Itemset Mining of Imaging Genetics GWAS in Alzheimer's Disease. *Genes (Basel)*". ( 2022) Jan 19;13(2):176. doi: 10.3390/genes13020176. PMID: 35205221; PMCID: PMC8871801.
- [14] Mutalib, Sofianita & Mohamed, Azlinah & Rahman, Shuzlina. "A Study on Frequent Itemset Mining for Identifying Associated Multiple SNPs". *Journal of Computer Science & Computational Mathematics*. (2019). 1-6. 10.20967/jcscm.2019.01.001.

# Speech Emotion through Voice & Accent

Kundan Sai Kotta, Sai Nikhil Samineni, Asst. Prof.G. Kavitha

Dept. of AI&MLCBIT(A)  
Hyderabad ,AP,India

**Abstract-** Detecting emotions through voice represents the next evolutionary leap in human-computer interaction, propelling us toward a more intuitive interface and enabling the development of superior recommendation systems. Voice, encompassing pitch, tone, and cadence, and accent, involving pronunciation patterns and linguistic nuances, play crucial roles in this context. Emotions, fundamental to human interaction, greatly influence communication and understanding. This research aims to investigate how variations in voice and accent contribute to expressing and interpreting emotions in speech. The study explores deep learning architectures and methodologies for this purpose, addressing associated challenges, limitations, and ethical considerations. Understanding the interplay of voice, accent, and emotions is pivotal for advancing technology in a beneficial manner.

**Index Terms-**voice, accent, emotion, intonation, deep learning.

## I. INTRODUCTION

In general existing systems, be it advanced chatbots or language models, primarily emphasize the conversion of spoken words to text, neglecting the crucial layer of emotional nuances that significantly impact human communication. While the focus on accurate transcription is vital, understanding and interpreting the emotional content embedded within speech adds an invaluable dimension to human-computer interaction (HCI). Recognizing the emotional state conveyed through speech is pivotal for creating empathetic and contextually aware systems that can respond appropriately to users. Emotion plays a vital role in a human being's life. The requirement for human-to-computer communication had become unavoidable. To accomplish this, a computer would have to respond differently based on how it perceives the scenario in the present. To make human-computer interaction more natural, the computer must respond to human emotions in the same way as people in similar situations would. To achieve the goal, the computer can identify emotion through facial expressions or voices. Speech is a significant technique of recognizing emotions in HCI. SER has become one of the most important aspects of HCI[6].

To achieve this, we inherently face two major limitations with the available research. At present, researchers predominantly employ deep neural networks to train machine learning models for emotion classification. This approach offers notable advantages including rapid training, high classification accuracy, and enhanced capability compared to traditional machine learning methods [5]. Conversely, traditional machine learning techniques often encounter challenges such as local optimization issues and limited generalization ability[3]. In this research we try to establish a model that is capable of properly classifying emotions as set a standard of emotional state. When the speaker must repress emotions, some parts of internal sensation are buried and are

not audible in speech. Therefore, computer-based systems are limited to what can be seen from the input of speech samples [6]. As a result of the lengthy dispute over the definition of "emotion" and the appropriate emotional classes, classifying emotional speech samples is a difficult task. To avoid that "fruitless discussion," Batliner et al. [7] favor the idea of emotion-related states [4]. However, among the systems that do acknowledge the importance of emotion detection, there remains a significant gap in addressing the diverse array of accents prevalent in the global population. Emotions are conveyed not only through words but also through variations in pitch, tone, and accent, which are unique to each individual and their cultural or regional background. Unfortunately, current emotion detection systems often fall short in effectively capturing these nuanced variations across accents.

But emotion detection from speech is quite difficult for many reasons: identifying the relevant emotion from a raw speech signal captured via a microphone can be affected by several factors such as gender, age, culture, health state, noise...The early Automatic Speech Recognition systems mainly focused on emotion recognition in several languages such as English[2]. Despite all the efforts, there has been little progress in determining which features were to select for improved performance [5].

Using a high-dimensional feature set that includes all sound parameters can aid in capturing all variances [6] but it can also lead to overfitting. A question could arise that whether a person expresses an emotion is largely dependent upon the person speaking, their culture, and the environment in which the person has been living. Majority of the study has concentrated on monolingual emotion classification, assuming that no cultural differences between speakers were present. Furthermore, application of large-scale acoustic parameters stood as a difficult task [7]. As a result of this, deep learning techniques are required for feature selection and low-latency SER. Through this study, we emphasize the significance of

not only recognizing emotions conveyed through words but also through variations in pitch, tone, and accent. Our

approach seeks to bridge this gap, paving the way for more inclusive emotion detection systems. By employing deep learning techniques for feature selection and low-latency Speech Emotion Recognition (SER), we aim to develop a model that accurately captures emotional nuances across various linguistic and cultural backgrounds, ultimately advancing the field of Human-Computer Interaction (HCI).

## II. LITERATURE SURVEY

- Dr. A. Arul Edwin Raj and Karan Kumar B have proposed a system where Mel-frequency Cepstral Coefficient (MFCC) feature is utilized to classify the data into different emotion groups. CNN is widely used for pattern recognition due to its many features like Mel Frequency Cepstral Coefficients (MFCC), a relatively simple structure, and fewer parameters for model training, making it ideal for SER. This technique effectively achieves a suitable compromise between the real-time process's performance precision and computing volume. As a machine learning model, the Speech Emotion Recognition (SER) system was developed.
- This research represents a new case study, aiming to construct and analyze an emotional speech corpus of the Algerian dialect. The objective is to propose a novel hybrid classification model designed to recognize emotions from Arabic speech. In pursuit of the research goals, a substantial annotated dataset comprising 1202 audio records was meticulously collected and constructed. These recordings were annotated with emotional labels such as happy, angry, neutral, or sad. Several experiments were conducted utilizing a variety of machine learning classification algorithms, in addition to deep convolutional and recurrent neural networks. It was observed that our proposed LSTM-CNN model surpassed all other classifiers and approaches, achieving an impressive accuracy of 93.34%. These results underscore the potential of LSTM networks in yielding compelling outcomes for speech emotion recognition. This is particularly significant in our case study, focusing on emotion detection from the Algerian dialect.
- Deep learning, a novel form of unsupervised methodology, employs artificial neural network models to analyze and process emotional information within the data. It utilizes data features in the deep learning process to identify and label the emotional content, thus enhancing understanding and analysis of the knowledge embedded within the dataset. This paper proposes a hybrid semantic text feature, integrating CNN and machine learning algorithms. The CNN algorithm is employed to analyze the data features, contributing to the development of a comprehensive and stable training model. Through human network training, the model is fine-tuned using CNN to label emotional content, and the data features generated during the deep learning process are utilized to configure the design parameters of the neural network model.
- Their custom feedforward-based deep learning model for speech emotion recognition demonstrated an impressive test accuracy of 93% and a training accuracy of 97.44%. Additionally, the test loss was recorded at 0.20, while the training loss stood at 0.081. Figure 6 in their study depicted the training and testing loss, along with the accuracy, illustrating a favorable performance during testing. The tools employed for simulation encompassed Python, Tensorflow, Numpy, Keras, and Google ColabPro. Their research primarily revolved around leveraging identifiers present in the datasets, including modality, vocal channel, emotional intensity, statement, repetition, and actor. These identifiers closely resembled stimulus characteristics. Notably, each expression was produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. The overarching goal was to accurately label these emotions following a thorough assessment of the audio files. While the achieved results were commendable, they acknowledged potential areas for improvement. These areas included implementing noise filtering and conducting a more detailed analysis of the language used. They recognized that accurately determining the correct labels based solely on the provided identifiers could be challenging at times.
- In this experiment, the researchers applied speech emotion recognition to the IEMOCAP dataset. They utilized spectral time-frequency information as a feature extraction technology, incorporating various operations and filtering processes. For classification, deep learning technology was employed, utilizing a CNN model to effectively capture advanced features that preserve emotional characteristics in speech. Additionally, an LSTM model was used to maintain temporal information characteristics in speech. During the test stage, the Weighted Accuracy (WA) achieved a rate of 61%, and the Unweighted Accuracy (UA) stood at 56%. In their future studies, the researchers intend to shift their focus towards addressing neutral confusion and non-neutral confusion. They believe that resolving the neutral confusion problem could significantly enhance the recognition rate for emotion recognition.
- The application successfully achieved speech emotion recognition and underwent comprehensive testing for both functional and non-functional requirements, employing an ELM and RF classifier combination as its engine. The SER application met all 6 out of 6 defined requirements with 100% accuracy during the identification of 70 speech data instances.
- In the proposed method, CNN emerged as the superior and more reliable choice compared to previous architectures. The Speech Emotion Recognition (SER) model developed through this approach holds considerable significance in practical scenarios like police stations, car board systems, and call centers, where accurately interpreting emotions is vital. However, the

study encountered certain limitations, notably the challenge of handling large-scale acoustic parameters.

Additionally, the research primarily focused on monolingual emotion classification, assuming a homogeneity of cultural context among speakers. In acknowledging these limitations and aiming to enhance the model's versatility and accuracy, future research endeavors should address these concerns. Exploring various languages, such as Korean and Spanish, in the SER model is essential for broader applicability. Moreover, integrating LSTM and other deep learning models could potentially augment the precision and efficiency of the proposed method.

- The research team utilized Visual Studio to leverage obtained output for emotion detection using a trained model classifier, presenting the results accordingly. The study encompassed a comprehensive analysis of training and testing samples, along with the identified features. Accuracy and confusion matrix were derived by comparing predictions post-training showcasing an accuracy of 81.52% in the confusion matrix. This result indicates the model's effectiveness, with efforts underway to further improve prediction accuracy by extracting additional features. The research findings underscore the significance of this technique in the scientific and technical domain. Librosa was a key tool utilized to extract emotion recognition features, while Pyaudio facilitated audio recording. The Matplotlib module played a vital role in visualizing audio waves for future reference. Employing a classifier model, the team successfully categorized various emotions, setting a promising foundation for future advancements in this domain.

### III. METHODOLOGY

The primary objective is to systematically collect, curate, and annotate new data tailored to suit identified algorithms for constructing the model. Extensive study of the data's features and their implications has informed the selection of specific models. Additionally, a thorough examination of the Evaluation Metrics outlined in various research papers has been conducted, with careful consideration given to integrating them for the robust evaluation of our model.

#### 1. Data Collection

In the domain of data collection, our focal points centered on two critical aspects:

- Emotion-based Audio File Classification
- Incorporating Diverse Accent Variations within the Audio Dataset

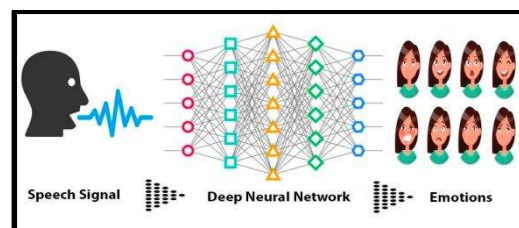
All datasets consist of audio files. The audio files are processed by converting them into appropriate features, such as Mel-frequency cepstral coefficients (MFCCs), using relevant techniques. Necessary data pre processing, including augmentation and normalization, is performed using Keras. A comprehensive model architecture is proposed, incorporating Convolutional Neural Networks (CNN) for feature extraction

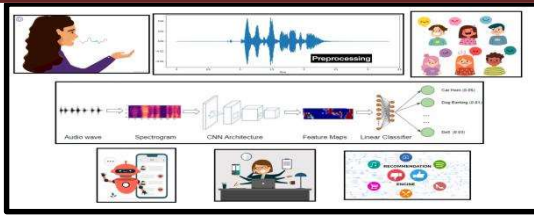
from Mel-frequency cepstral coefficients (MFCCs) and Recurrent Neural Networks (RNN) for capturing sequential patterns. The model undergoes training on the curated dataset for a specified number of epochs. Evaluation involves testing the model with various audio inputs, including diverse accents, to assess its performance and achieve higher accuracy in voice emotion detection. The trained model is saved for future use, and a web application is developed using the Flask framework, enabling users to upload audio files for voice emotion analysis.

#### 2. Speech Emotion Detection Process and Datasets

Analyzing speech emotion through voice and accent in real-time involves a structured process. The first step is data collection, gathering a diverse dataset of audio recordings encompassing various emotions and accents, utilizing platforms like Kaggle or recording custom data. Next, audio preprocessing is performed, extracting features like MFCC, pitch, and intensity analysis, and segmenting the audio into manageable units. Accent analysis is then implemented, utilizing linguistic analysis and pronunciation patterns to categorize accents. The subsequent step involves model selection and building, choosing appropriate deep learning models like CNNs, RNNs, or hybrids, and designing an architecture that integrates both emotion recognition and accent analysis aspects.

Afterward, the model is trained by splitting the dataset into training and testing sets, ensuring diverse representation, and using the chosen architecture and features. Integration into a real-time system follows, allowing live audio input or audio file processing. Real-time inference is then implemented, processing the audio input and providing predictions for both emotion and accent. Evaluation involves assessing model performance using metrics like accuracy, precision, recall, F1 score, and confusion matrix for both emotion and accent recognition, leading to fine-tuning based on results. Finally, integration into applications like customer service, virtual assistants, or mental health analysis is conducted, enabling a comprehensive system to recognize emotions and accents in real-time audio data, providing valuable insights and enhancing communication in various applications.





**Datasets**

Dataset Link	Locality
<a href="https://www.kaggle.com/datasets/uwrfkagglerravdess-emotional-speech-audio">https://www.kaggle.com/datasets/uwrfkagglerravdess-emotional-speech-audio</a>	Well documented, clean dataset for UK dialect.
<a href="https://www.kaggle.com/code/lkergalipatak/speech-emotion-recognition-with-cnn/input?select=Crema">https://www.kaggle.com/code/lkergalipatak/speech-emotion-recognition-with-cnn/input?select=Crema</a>	Augmented US dialect data.
<a href="https://www.kaggle.com/datasets/tapakah68/emotions-on-audio-dataset">https://www.kaggle.com/datasets/tapakah68/emotions-on-audio-dataset</a>	GLOBAL
<a href="https://www.openu.ac.il/home/hassner">https://www.openu.ac.il/home/hassner</a>	GLOBAL
<a href="https://rock.github.io/XD-Violence">https://rock.github.io/XD-Violence</a>	GLOBAL
<a href="http://en.arabicspeechcorpus.com/">http://en.arabicspeechcorpus.com/</a>	ARABIC Phd Collection
<a href="https://gitlab.com/nicolasobin/att-hack/-/blob/master/README.md">https://gitlab.com/nicolasobin/att-hack/-/blob/master/README.md</a>	FRENCH (Data )

**3. Emotion detection using Deep Learning techniques**

Deep learning has emerged as a powerful tool in recent years for predicting crime, leveraging various algorithms such as Convolutional Neural Networks (CNN), sentiment analysis, and deep neural networks. These algorithms are proficient in detecting patterns and anomalies in data, including text, images, audio, and social media, and can provide insights into potential criminal activities. When adapted for speech emotion detection through voice and accent analysis, these algorithms can prove instrumental.

- Customized CNN: CNN (Convolutional Neural Network) is a well-known algorithm used not only in image processing but also in speech emotion classification. In the context of analyzing audio data, the CNN model is customized to handle audio features extracted from speech signals. By assigning weights and biases to differentiate features relevant to various emotions, this CNN variant with 32, 64, and 128 filters effectively captures the emotional characteristics embedded in speech data. The model performs optimally on both training and testing datasets, highlighting its potential in speech emotion analysis.

- R-CNN: The concept of Region-based Convolutional Neural Network (R-CNN) is extended to speech emotion recognition, demonstrating its versatility beyond image analysis. By employing R-CNN architecture, features from different regions of the audio signal are extracted, enabling effective emotion recognition. The model, utilizing 32 filters and subsequent layers, can efficiently analyze audio data, providing valuable insights into the emotional content of speech. The integration of R-CNN in speech emotion recognition signifies a promising advancement in understanding and interpreting emotions through voice.
- VGGNET19 Adaptation: Adapting the VGGNET19 architecture, renowned for image classification, to the domain of speech emotion analysis showcases its applicability in various data types. In this context, VGGNET19 is tailored to handle audio features extracted from speech signals. The model, employing convolution layers and max-pooling, effectively processes the audio data, recognizing patterns that signify different emotions. This adaptation underscores the adaptability and effectiveness of deep learning architectures in analyzing emotional content in speech.
- ResNet50+LSTM Fusion: Combining Residual Network (ResNet) with Long Short-Term Memory (LSTM) networks presents a powerful approach for speech emotion classification. ResNet, known for its depth and accuracy, is utilized in feature extraction from audio data. The LSTM network, designed to analyze sequential data, effectively captures temporal dependencies in speech, enhancing emotion classification accuracy. The fusion of ResNet50 and LSTM offers a compelling solution for analyzing emotional nuances conveyed through speech signals.
- YOLOv5-inspired Approach : Inspired by the efficiency of YOLOv5 (You Only Look Once) in object detection, a similar approach is applied to real-time speech emotion recognition. The YOLOv5 architecture, designed for speed and accuracy, is adapted to efficiently process audio features extracted from speech signals. This innovative approach enables real-time analysis of emotional content in speech, showcasing the adaptability of cutting-edge techniques for audio-based applications.
- Simplified YOLO: The essence of the You Only Look Once (YOLO) algorithm, initially developed for object detection, is harnessed to efficiently detect and classify emotions in speech. This simplified YOLO model focuses on analyzing audio features encompassing emotional cues. By treating speech emotion detection as a regression problem, this approach combines efficiency with acceptable accuracy, showcasing its potential in real-time emotion analysis through speech.
- MobileNet: MobileNet, a lightweight deep learning architecture, is repurposed for speech emotion recognition. Its efficiency and computational speed make it an ideal choice for analyzing audio features extracted from speech. By processing audio data with precision and speed, MobileNet proves to be a valuable tool for real-time emotion

recognition in speech, paving the way for lightweight yet effective emotion analysis applications.

- Xception: Xception, known for its accuracy and efficiency in image recognition, is fine-tuned to handle audio features in speech emotion recognition. By leveraging its depth-wise separable convolutional network, Xception efficiently processes audio data, recognizing patterns indicative of various emotions. The adaptability of Xception in analyzing

audio signals underscores its potential in enhancing speech emotion recognition accuracy.

- InceptionV3+LSTM: InceptionV3, recognized for its prowess in image recognition, is combined with LSTM to capture emotional patterns in sequential speech data. InceptionV3 extracts hierarchical features from audio signals, while LSTM analyzes sequential data to detect emotional nuances over time. This fusion provides a comprehensive approach to speech emotion recognition, offering insights into emotional variations within speech sequences.
- VGG16+LSTM: VGG16, a renowned deep CNN architecture, is paired with LSTM to track emotional dynamics within speech sequences. VGG16 effectively locates and tracks emotional cues, while LSTM identifies anomalies and patterns in speech sequences. This combined approach proves valuable in understanding how emotions evolve and manifest in speech over time, enhancing the accuracy of emotion classification in audio data.

#### IV.CONCLUSION

In striving to enhance human-computer interaction and accentuate the role of pitch and accent in communication, this research presents a method for real-time analysis of voice emotions. With a focus on achieving high training accuracy and minimizing loss during training, we propose a real-time system for emotion detection utilizing CNNs. This system adeptly discerns a range of emotions, even amidst diverse accents and varying pitch, enriching the user experience in human-computer interaction.

#### REFERENCES

1. Arul Edwin Raj, K. K. B, S. S and R. A, "Speech Emotion Recognition using Deep Learning," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 505-509, doi: 10.1109/ICIDCA56705.2023.10100056.
2. R. Y. Cherif, A. Moussaoui, N. Frahta and M. Berrimi, "Effective speech emotion recognition using deep learning approaches for Algerian dialect," 2021 International Conference of Women in Data Science at Taif University (WiDSTaif ), Taif, Saudi Arabia, 2021, pp. 1-6, doi: 10.1109/WiDSTaif52235.2021.9430224.

3. W. Wang, G. Wen and Z. Zheng, "Design of Deep Learning Mixed Language Short Text Sentiment Classification System Based on CNN Algorithm," 2022 IEEE 2nd International Conference on Mobile Networks and Wireless Communications (ICMNWC), Tumkur, Karnataka, India, 2022, pp. 1-5, doi: 10.1109/ICMNWC56175.2022.10031786.

4. D. Femi and S. Thylashri, "Human Voice Emotion Recognition Using Multilayer Perceptron," 2022

International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICES), Chennai, India, 2022, pp. 1-4, doi: 10.1109/ICES55317.2022.9914336.

5. K. -Y. Huang, C. -H. Wu, Q. -B. Hong, M. -H. Su and Y. -H. Chen, "Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 5866-5870, doi: 10.1109/ICASSP.2019.8682283.

6. M. Saloumi et al., "Speech Emotion Recognition Using One-Dimensional Convolutional Neural Networks," 2023 46th International Conference on Telecommunications and Signal Processing (TSP), Prague, Czech Republic, 2023, pp. 212-216, doi: 10.1109/TSP59544.2023.10197766.

7. Ainurrochman, I. I. Febriansyah and U. L. Yuhana, "SER: Speech Emotion Recognition Application Based on Extreme Learning Machine," 2021 13th International Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia, 2021, pp. 179-183, doi: 10.1109/ICTS52701.2021.9609016.

8. U. Mahesh YadavKonangi, V. R. Katreddy, S. K. Rasula, G. Marisa and T. Thakur, "Emotion Recognition through Speech: A Review," 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2022, pp. 1150-1153, doi: 10.1109/ICAAIC53929.2022.9792710.

9. H. Li, X. Zhang and M. -J. Wang, "Research on Speech Emotion Recognition Based on Deep Neural Network," 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 2021, pp. 795-799, doi: 10.1109/ICSIP52628.2021.9689043.

10. K. -Y. Huang, C. -H. Wu, Q. -B. Hong, M. -H. Su and Y. -H. Chen, "Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Brighton, UK, 2019, pp. 5866-5870, doi: 10.1109/ICASSP.2019.8682283.



ISSN:2147-6799

# International Journal of INTELLIGENT SYSTEMS AND APPLIED ENGINEERING

www.ijisae.org

## Predicting Pedestrian Behavior at Zebra Crossings using Pose Estimation and Deep Learning

Pannalal Boda<sup>1</sup>, Y. Ramadevi<sup>2</sup>

**Submitted:** 09/09/2023

**Revised:** 21/10/2023

**Accepted:** 06/11/2023

**Abstract:** Anticipating pedestrian behavior is critical for traffic management, developing Advanced Driver Assistance Systems (ADAS), and creating autonomous vehicles. However, the unpredictability of pedestrians at zebra crossings poses a challenge in designing systems that can aid drivers or enable self-driving. Existing studies often overlook pedestrian behavior, focusing on predicting motion, and there is no integrated system that connects perception and decision-making tasks. This paper proposes a new bottom-up pedestrian Pose Estimation model based on a CNN network that is trained on a Pretrained model. This model allows for the analysis of videos captured at zebra crossings and enables the detection of pedestrian poses and movements such as walking, standing, hand signals, crossing, and not crossing. The model is evaluated on the pedestrian intention estimation (PIE) dataset using the COCO-18 key point model. Our approach provides a solution for predicting pedestrian behavior at zebra crossings. Machine learning-based classifiers are used to evaluate the performance across different prediction horizon values, resulting in improved accuracy and efficiency. The study has significant implications for traffic management, ADAS, and autonomous vehicles, as it enables them to better understand and predict pedestrian actions. Overall, this study highlights the importance of integrating perception and decision-making in predicting pedestrian behavior and provides a promising solution for addressing this critical problem.

**Keywords:** *Pedestrians' pose estimation, behavioral analysis, Advanced Driver Assistance Systems, Behavior classification.*

### 1. Introduction

According to the World Health Organization, road traffic accidents cause 1.35 million deaths each year, with vulnerable road users, including pedestrians [1], accounting for more than half of those killed. With the increasing prevalence of connected autonomous vehicles (CAVs) [2], protecting pedestrians has become even more critical. Predicting the behavior of pedestrians in zebra crossing zones is essential for autonomous vehicle navigation, but it is challenging because pedestrians do

computers must understand the context, accuracy is crucial. The goal is to recognize pedestrian behavior. The number of times people cross the street. Pedestrian detection is an essential function of autonomous vehicles. Developing the ability to recognize pedestrian behavior is equally crucial.

To address this challenge, this paper uses a dataset of videos and posture estimation to detect important landmarks on the body.



ISSN:2147-6799

# International Journal of INTELLIGENT SYSTEMS AND APPLIC ENGINEERING

www.ijisae.org

## Binary Image Classification on Fashion-MNIST Using Quantum and CIRQ

Prabhakar Kandukuri<sup>1</sup>, Dasari N. V. Syam Kumar<sup>2</sup>, V. Sessa Srinivas<sup>3</sup>, Chiluka  
Kranthi Kumar Singamaneni<sup>5\*</sup>

Submitted: 27/08/2023

Revised: 22/10/2023

Accepted: 02/11/2023

**Abstract:** TensorFlow and Cirq, two key Google frameworks, are used to process the binary image classification. TensorFlow and Cirq frameworks were developed by Google. The binary image classification is utilized most frequently in object from its background. The process of segmentation makes it possible to name each pixel as either assign black and white colors that match to those labels. The combination of machine learning with quantum classification that is superior to that achieved by machine learning classification techniques. The TensorFlow quantum machine learning framework that enables quick prototyping of hybrid quantum-classical ML models using the TFQ library. In order to process the categorization, QNN and CNN are both used as algorithms. Existing classification include overfitting, a limited amount of data, variability in picture data, and background interrelated. The quantum machine learning methodology that has been developed has the potential to recognize image data, optimize the background noise that has been discovered in the images, and minimize the overfitting data.

**Keywords:** TensorFlow-Quantum, CNN, QNN, Cirq

### 1. Introduction

Binary image classification is a popular task in machine learning where the goal is to classify images into two distinct classes. The Fashion MNIST dataset, which contains images of clothing items, is a popular benchmark dataset for binary image classification tasks. Recently, researchers have been exploring the usage of quantum computing techniques to progress the accuracy and speed of image classification tasks. TensorFlow-Quantum (TFQ) is a library developed by Google that allows the integration of quantum computing into TensorFlow, a popular machine-learning framework. TFQ offers a great level API for

By combining the power of TensorFlow and Cirq, researchers can build and train quantum machine learning models for binary image classification on the Fashion MNIST dataset. These models have the potential for higher accuracy and faster training times compared to classical machine learning models, leading to improved performance in image classification and other related tasks. Several quantum machine learning algorithms have been proposed to identify the optimal parameters for the classical and the quantum parts of the model [3][5] [12-13]. The quantum machine learning models that are used like TensorFlow-Quantum and Cirq are producing quite optimal results.





ISSN:2147-6799

# International Journal of INTELLIGENT SYSTEMS AND APPLIED ENGINEERING

www.ijisae.org

## Deciphering Market Dynamics: A Data Science and Machine Learning Approach Using Chaos Theory for Trend Prediction

<sup>1</sup>K. Prabhakar\*, <sup>2</sup>Manjula V., <sup>3</sup>P. Punitha, <sup>4</sup>Khasim Vali Dudekula, <sup>5</sup>Panduranga

Submitted: 11/10/2023

Revised: 30/11/2023

Accepted: 10/12/2023

**Abstract:** This study introduces an innovative technique for the prediction of financial market movements using chaos theory principles. Employing time-delay embedding alongside attractor reconstruction, the study identifies complex structures within financial market time series data. The identification of these patterns facilitates the development of a predictive model aimed at forecasting forthcoming market behaviours. The findings of the research challenge the traditional view of financial markets as random; however, the application of chaos theory offers a valuable perspective into the intricate mechanisms governing these sophisticated systems. The study highlights the potential of chaos theory as a tool in deciphering and anticipating the fluctuations contributing to the fields of economic forecasting and financial analysis.

**Index Terms**—Chaos Theory, Time Delay Embedding, Attractor Reconstruction, Trend Prediction, Financial Market Analysis

### I. Introduction

In the ever-evolving landscape of financial markets, the ability to predict market trends is a coveted asset for economists, traders, and financial analysts alike. Traditional models often fall short in capturing the complexities and dynamic nature of financial systems. This limitation has prompted a pursuit for alternative approaches that can accommodate the inherent unpredictability and non-linear characteristics of economic data. Enter chaos theory – a framework initially developed to understand complex natural systems, now increasingly relevant in the domain of

financial market analysis. Chaos theory reveals the underlying patterns, complexities, and self-similarity in financial data, offering a unique lens through which trends might be anticipated. Drawing inspiration from this novel methodology that integrates time delay embedding and attractor reconstruction, this study identifies significant patterns in financial market data. The introduction of this methodology provides a nuanced understanding of non-linear market dynamics. This paper explores the intricacies of this approach, its application in forecasting future market

*1 Professor, Dept. of AIML, Chaitanya Bharathi Institute of Technology, Gandhinagar, Hyderabad, India*

# Enhancing IoT Network Security: ML and Blockchain for Intrusion Detection

N. Sunanda<sup>1</sup>, K. Shailaja<sup>2</sup>, Prabhakar Kandukuri<sup>3</sup>,

Krishnamoorthy<sup>4</sup>, Vuda Sreenivasa Rao<sup>5</sup>, Sanjiv Rao Godla<sup>6</sup>

Assistant Professor, Department of CSE-(CyS,DS) and AI&DS ,VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India<sup>1</sup>

Associate Professor, Department of CSE, Vasavi College of Engineering, Hyderabad, India<sup>2</sup>

Professor, Department of Artificial Intelligence and Machine Learning,

Chaitanya Bharathi Institute of Technology - Hyderabad, India<sup>3</sup>

Associate Professor, Department of CSE, Panimalar Engineering College, Chennai, India<sup>4</sup>

Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India<sup>5</sup>

Professor, Department of CSE (Artificial Intelligence & Machine Learning), Aditya College of Engineering & Technology - Surampalem, Andhra Pradesh, India<sup>6</sup>

**Abstract**—Given the proliferation of connected devices and the evolving threat landscape, intrusion detection plays a pivotal role in safeguarding IoT networks. However, traditional methodologies struggle to adapt to the dynamic and diverse settings of IoT environments. To address these challenges, this study proposes an innovative framework that leverages machine learning, specifically Red Fox Optimization (RFO) for feature selection, and Attention-based Bidirectional Long Short-Term Memory (Bi-LSTM). Additionally, the integration of blockchain technology is explored to provide immutable and tamper-proof logs of detected intrusions, bolstering the overall security of the system. Previous research has highlighted the limitations of conventional intrusion detection techniques in IoT networks, particularly in accommodating diverse data sources and rapidly evolving attack strategies. The attention mechanism enables the model to concentrate on pertinent features, enhancing the accuracy and efficiency of anomaly and malicious activity detection in IoT traffic. Furthermore, the utilization of RFO for feature selection aims to reduce data dimensionality and enhance the scalability of the intrusion detection system. Moreover, the inclusion of blockchain technology enhances security by ensuring the integrity and immutability of intrusion detection logs. The proposed framework is implemented using Python for machine learning tasks and Solidity for blockchain development. Experimental findings demonstrate the efficacy of the approach, achieving a detection accuracy of approximately 98.9% on real-world IoT datasets. These results underscore the significance of the research in advancing IoT security practices. By amalgamating machine learning, optimization techniques, and blockchain technology, this framework provides a robust and scalable solution for intrusion detection in IoT networks, fostering improved efficiency and security in interconnected environments.

**Keywords**—Intrusion detection; IoT networks; machine learning; random forest, red fox optimization; blockchain technology

## I. INTRODUCTION

The Internet of Things (IoT) represents a transformative innovation in automation and connectivity, comprising a vast network of interconnected devices equipped with actuators, sensors, and computational capabilities [1]. These devices encompass a diverse range, from everyday items like household appliances and wearables to complex industrial machinery and infrastructure components. Central to IoT networks is their autonomous ability to collect, process, and transmit data, eliminating the need for direct human intervention. This autonomy empowers organizations and individuals to leverage data-driven insights and automation across various sectors and industries. For instance, in smart homes, IoT devices facilitate energy monitoring, remote appliance control, and enhanced security via connected surveillance systems [2].

Wearable sensors and medical gadgets help with early health issue diagnosis, individualized treatment strategies, and remote patient monitoring in the healthcare industry. In transportation, IoT technologies optimize logistics, improve traffic management, and enhance passenger safety through intelligent vehicle systems and infrastructure. Moreover, IoT networks extend their reach into diverse sectors such as agriculture, where precision farming techniques leverage sensor data to optimize irrigation, monitor soil conditions, and maximize crop yields[3]. In industrial settings, IoT-enabled machinery and production systems enable predictive maintenance, real-time monitoring of equipment health, and automation of manufacturing processes, leading to increased efficiency and reduced downtime. The overarching goal of IoT networks is to enhance connectivity, efficiency, and convenience while enabling new levels of automation and control across various domains. By seamlessly integrating physical devices with digital technologies, IoT networks pave the way for a more interconnected and intelligent world, where data-driven insights drive decision-making and innovation. However, this proliferation of connected devices

also brings about significant challenges, particularly in terms of security, privacy, and interoperability, which must be addressed to fully realize the potential benefits of the IoT revolution [4].

IoT networks exhibit a high degree of heterogeneity, encompassing a diverse array of devices with varying computational capabilities, communication protocols, and operating systems. From simple sensors to complex smart appliances and industrial machinery, these devices run on different platforms, including embedded systems, Linux-based platforms, and proprietary firmware[5]. This heterogeneity poses challenges for interoperability and standardization. Moreover, IoT networks are highly scalable, capable of supporting deployments ranging from small-scale implementations to massive infrastructures comprising millions of interconnected devices. This scalability leads to complex network topologies and management challenges. Connectivity serves as a cornerstone for IoT networks, with devices employing a range of wired and wireless communication technologies. The selection of connectivity technology is influenced by factors such as range, power consumption, and deployment environment. Additionally, IoT networks generate a wide array of data types, including sensor readings, images, audio, and video streams, presenting challenges for data processing and analysis. Effectively managing this data diversity is essential for deriving meaningful insights while maintaining scalability, efficiency, and data privacy [6].

IoT networks are susceptible to a myriad of security vulnerabilities, posing significant challenges to their integrity and reliability. Weak authentication and authorization mechanisms represent a prevalent threat, as many IoT devices are shipped with default or easily guessable credentials, providing malicious actors with unauthorized access and control over these devices [7]. Furthermore, insecure communication practices exacerbate the risk, as IoT devices often transmit data over unencrypted channels or employ weak encryption protocols, leaving sensitive information vulnerable to eavesdropping and interception by malicious entities. Compounding these issues is the lack of timely security updates from manufacturers, leaving devices exposed to known vulnerabilities and exploits. Physical vulnerabilities also pose a substantial risk to IoT networks, as attackers can exploit physical access to tamper with hardware components, extract sensitive data, or implant malicious firmware, compromising the integrity and functionality of these devices [8].

Additionally, IoT devices are susceptible to being co-opted into botnets and used to launch distributed denial-of-service (DoS) attacks against targeted services or networks, leading to disruptions and downtime. Moreover, the vast amounts of personal and sensitive data collected and transmitted by IoT devices raise significant privacy concerns, including unauthorized access, data breaches, and misuse of information. Supply chain risks further exacerbate the security landscape, as the global supply chain for IoT devices is often complex and opaque, making it challenging to verify the integrity and authenticity of hardware components and software firmware [9]. Lastly, interoperability issues between

IoT devices and protocols introduce additional vulnerabilities, enabling attackers to exploit weaknesses in communication interfaces and protocols, potentially compromising the entire network. A comprehensive strategy that includes strong authentication procedures, encryption methods, regular security upgrades, physical security measures, and privacy-enhancing technology is needed to address these issues. In addition, stakeholders need to work together to create industry-wide guidelines and recommendations for protecting IoT networks and devices, minimizing risks, and guaranteeing the dependability and trustworthiness of the IoT ecosystems [10].

Intrusion detection in IoT networks is hindered by the dynamic and heterogeneous nature of these environments, along with the continuously evolving threat landscape. Traditional methods struggle to adapt to the diverse array of devices, communication protocols, and data formats present in IoT networks, leading to limited coverage and effectiveness. Scalability poses another challenge, as the sheer volume of interconnected devices generates large amounts of data that traditional systems may struggle to process in real-time. Resource constraints on IoT devices further complicate matters, making it difficult to deploy traditional intrusion detection solutions. Furthermore, newer or undiscovered threats could not be detected by conventional techniques, calling for more sophisticated detection capabilities. Moreover, worries about data privacy and integrity continue since centralized systems have the potential to expose vulnerabilities or corrupt critical data. Innovative solutions that are suited to the special features of internet of things networks are needed to tackle these issues. These solutions must be scalable, resource-efficient, capable of robust detection, and equipped with improved security mechanisms to efficiently reduce hazards [11].

The rapid expansion of Internet of Things (IoT) networks has underscored the critical need for a robust and scalable intrusion detection framework capable of effectively mitigating security threats. Traditional intrusion detection systems (IDS) often struggle to adapt to the dynamic and heterogeneous nature of IoT environments, necessitating innovative solutions. Our research is motivated by the imperative to develop such a framework, leveraging advanced machine learning techniques like Attention-based Bidirectional Long Short-Term Memory (BiLSTM) networks for real-time threat detection. Additionally, the integration of Red Fox Optimization (RFO) enhances the efficiency of feature selection, enabling more accurate identification of relevant data amidst the complexities of IoT networks. Furthermore, the incorporation of blockchain technology ensures the integrity and trustworthiness of intrusion detection data, facilitating transparent incident response and forensic analysis. By synergizing these technologies, our framework offers a comprehensive defense mechanism against evolving threats, safeguarding critical assets and bolstering the security posture of IoT ecosystems. The key contribution of the research is stated as follows:

- The research presents a pioneering framework that combines machine learning techniques, such as Attention-based BiLSTM networks, with Red Fox

Optimization for feature selection, providing a novel approach to intrusion detection in IoT networks.

- By leveraging advanced machine learning algorithms, our framework achieves a significantly higher detection accuracy of approximately 98%, surpassing traditional intrusion detection systems and effectively mitigating security threats in IoT environments.
- The integration of Red Fox Optimization streamlines feature selection, enhancing the scalability and efficiency of our framework in handling the dynamic and heterogeneous nature of IoT data streams, thus ensuring robust performance even in large-scale IoT deployments.
- Incorporating blockchain technology ensures the integrity and tamper-resistance of intrusion detection data, providing transparent incident response and forensic analysis capabilities, thereby enhancing the overall security and trustworthiness of IoT networks.

The paper begins with an introduction to the research topic in Section I, followed by a comprehensive review of related literature in Section II. The methodology in Section IV outlines the proposed framework's design and implementation, with Section V covering experimental evaluation, results analysis, and discussion on the framework's effectiveness. Finally, Section VI concludes the paper.

## II. RELATED WORKS

Strong security mechanisms inside IoT networks are vital, as evidenced by the increasing ubiquity of Internet of Things (IoT) technologies. But in Internet of Things contexts, conventional intrusion detection systems face severe restrictions because of limited resources and the intrinsic complexity of the network. Liang et al. [12] research aims to tackle these issues by developing, putting into practice, and assessing a novel intrusion detection system. This system makes use of deep learning algorithms, blockchain technology, and multi-agent systems as part of a hybrid placement strategy. The data collecting, management, analysis, and reaction components of the system are organised into separate modules. The National Security Lab's NSL-KDD dataset was used for experimental verification, which demonstrates how well deep learning algorithms detect assaults, especially at the IoT network's transport layer. Notwithstanding the encouraging outcomes, the study admits significant limitations, such as the requirement for additional improvement and optimisation of the suggested system in order to guarantee its scalability and suitability for use in a variety of IoT scenarios.

Alkadi et al. [13] paper presents a novel approach to collaborative intrusion detection for safeguarding IoT and cloud networks, leveraging the capabilities of deep blockchain technology. By integrating blockchain into intrusion detection systems, the proposed framework aims to enhance the security posture of interconnected environments through collaborative threat intelligence sharing and consensus-driven decision-making processes. Through the utilization of machine learning algorithms and distributed ledger technology, the framework

enables real-time detection and response to emerging threats across diverse network landscapes. Experimental results demonstrate the efficacy of the framework in detecting intrusions and mitigating security risks in various network scenarios. However, the adoption of deep blockchain technology introduces challenges related to scalability, latency, and resource consumption. The computational overhead associated with maintaining a distributed ledger across multiple nodes may impact the real-time responsiveness of the intrusion detection system. Furthermore, ensuring consensus among distributed nodes in a timely manner can pose synchronization and coordination challenges, potentially affecting the system's overall efficiency and effectiveness in rapidly evolving threat landscapes. Addressing these scalability and performance limitations is essential to realize the full potential of the proposed framework in large-scale IoT and cloud networks.

The necessity for strong security measures to protect Internet-of-things (IoT) environments from potential threats has been highlighted by the growth of IoT devices. In order to protect computer networks, including the Internet of Things, from many types of security breaches, intrusion detection systems, or IDSs, are essential. The utilisation of collaborative intrusion detection systems or networks, also known as CIDSs or CIDNs, has shown promise in improving detection performance through the sharing of vital information across IDS nodes, including signatures and alarms. Nevertheless, because collaborative networks are distributed, they are vulnerable to insider assaults, in which rogue nodes spread fake signatures, jeopardising the accuracy and effectiveness of intrusion detection systems. Using blockchain technology presents a viable way to safely validate shared signatures. In this regard, the research of Li et al. (Li et al. 2019) presents CBSigIDS, an innovative framework for blockchain-based collaborative signature-based IDSs intended to create and gradually update a trusted signature database in collaborative IoT contexts. With no need for a reliable middleman, CBSigIDS provides a verified method in distributed architectures. Although CBSigIDS shows promise in strengthening the efficiency and robustness of signature-based IDSs, a significant disadvantage is the possible overhead related to blockchain activities, which calls for additional optimisation to guarantee scalability and efficacy in practical deployments.

Issues with privacy, security, and single points of failure in centralised storage structures still exist as the Internet of Things (IoT) gains pace, especially in crucial applications. By providing decentralised and secure data management, blockchain technology has emerged as a viable answer to these problems. There is a lot of potential for improving social and economic advantages when blockchain is integrated with IoT. But as the 2017 attack on a pool of miners has shown, blockchain-enabled Internet of Things (IoT) networks are vulnerable to Distributed Denial of Service (DDoS) attacks, underscoring the necessity of strong security protocols. Furthermore, for efficient analysis and decision-making, these applications' enormous data generation demands the use of sophisticated analytical tools like machine learning (ML). In order to address these issues, a unique solution is presented in

the paper by Kumar et al. [14]. This paper presents a distributed Intrusion Detection System (IDS) intended to detect distributed denial of service (DDoS) assaults targeting mining pools within Internet of Things networks, using fog computing and blockchain technology. Using Random Forest (RF) and an optimised gradient tree boosting system (XGBoost), both trained on dispersed fog nodes, the efficacy of the suggested IDS is evaluated. The BoT-IoT dataset, which covers recent assaults seen in IoT networks with blockchain support, is used in the evaluation. The possible costs and difficulties of implementing a distributed IDS employing fog computing in practical settings might be a drawback of the recommended strategy, necessitating more study and optimisation for efficiency and scalability. However, the outcomes demonstrate that Random Forest outperforms XGBoost in multi-attack recognition and binary attack detection.

Protecting industrial IoT (IIoT) networks from security threats is crucial as these networks grow to be essential parts of vital infrastructure. Numerous strategies utilizing Blockchain algorithms and machine learning techniques have been investigated separately to overcome this problem. However, Vargas et al. [15] offer an integrated strategy in this research that integrates these approaches to produce a thorough defense mechanism for networks of Internet of Things devices. The objectives of this mechanism are to identify potential dangers, initiate safe channels for information exchange, and adjust to the processing power of industrial Internet of things settings. The suggested method offers a workable way to identify and stop intrusions in Internet of Things networks and shows effectiveness in accomplishing its goals. Despite its achievements, it's crucial to remember that the suggested integrated strategy can present challenges for management and implementation, necessitating the need for extra funding and knowledge for deployment in actual IIoT scenarios. More investigation is required to ensure scalability and efficiency while minimizing overhead by streamlining and optimizing the integration process.

### III. PROBLEM STATEMENT

Despite the notable advancements in intrusion detection systems (IDS) and the integration of blockchain technology and machine learning techniques in securing Internet of Things (IoT) networks, several research gaps persist. Existing studies focus predominantly on individual aspects such as deep learning algorithms, blockchain-based intrusion detection, or collaborative signature-based IDSs. However, there is a scarcity of research that comprehensively addresses the complex security challenges of IoT environments by integrating multiple technologies and methodologies. Furthermore, scalability, efficiency, and practical feasibility remain critical concerns across these studies, indicating the need for further exploration and refinement. Thus, our research aims to bridge this gap by proposing a holistic framework that combines deep learning algorithms, blockchain technology, and collaborative intrusion detection

mechanisms to provide robust security solutions for IoT networks. By addressing these multifaceted challenges and evaluating the proposed framework's scalability and effectiveness across diverse IoT scenarios, our research endeavors to contribute towards the development of comprehensive and practical security solutions tailored for IoT environments.

### IV. METHODOLOGICAL INTEGRATION OF ML AND BLOCKCHAIN FOR IOT INTRUSION DETECTION

The suggested method builds a strong intrusion detection system (IDS) that is suited for the complex architecture of Internet of Things networks by fusing blockchain technology with machine learning. Network traffic, sensor readings, device logs, and other data from IoT devices are first gathered and preprocessed to extract pertinent attributes that are essential for intrusion detection. The framework optimizes feature subsets to increase intrusion detection efficacy and efficiency using the Red Fox Optimization (RFO) approach. Then, real-time anomaly detection is achieved by using Attention (BiLSTM) networks, which take advantage of their capacity to process sequential data streams present in Internet of Things settings. Blockchain technology is easily incorporated to guarantee the immutability and integrity of intrusion detection data. Smart contracts are utilized to provide safe communication and consensus building across dispersed Internet of Things devices, guaranteeing the accuracy and consistency of the data. Benchmark datasets such as the NSL-KDD dataset are used to evaluate the framework's performance in detail across a range of intrusion situations. By employing this technique, researchers want to enhance the efficacy and security of intrusion detection in internet of things networks, as well as tackle the constantly evolving problems associated with IoT setups [16]. The suggested technique's architecture is depicted in Fig. 1.

#### A. Data Collection

The data collection process involves gathering information from IoT devices, drawing upon a diverse array of network traffic, sensor readings, and device logs. In this research, we utilize the NSL-KDD dataset, an open-source resource available on Kaggle [17], to facilitate the collection of comprehensive data for intrusion detection system development. The NSL-KDD dataset offers a rich repository of labeled network traffic data, encompassing various types of attacks and normal behaviors, thereby enabling thorough analysis and evaluation of intrusion detection algorithms. Leveraging this openly accessible dataset ensures transparency and reproducibility in our research methodology, allowing for robust validation and benchmarking of the proposed intrusion detection framework against a standardized dataset. Through meticulous data collection from the NSL-KDD dataset, we aim to capture the diverse range of potential threats and normal activities prevalent in IoT networks, laying the foundation for effective intrusion detection system design and evaluation.

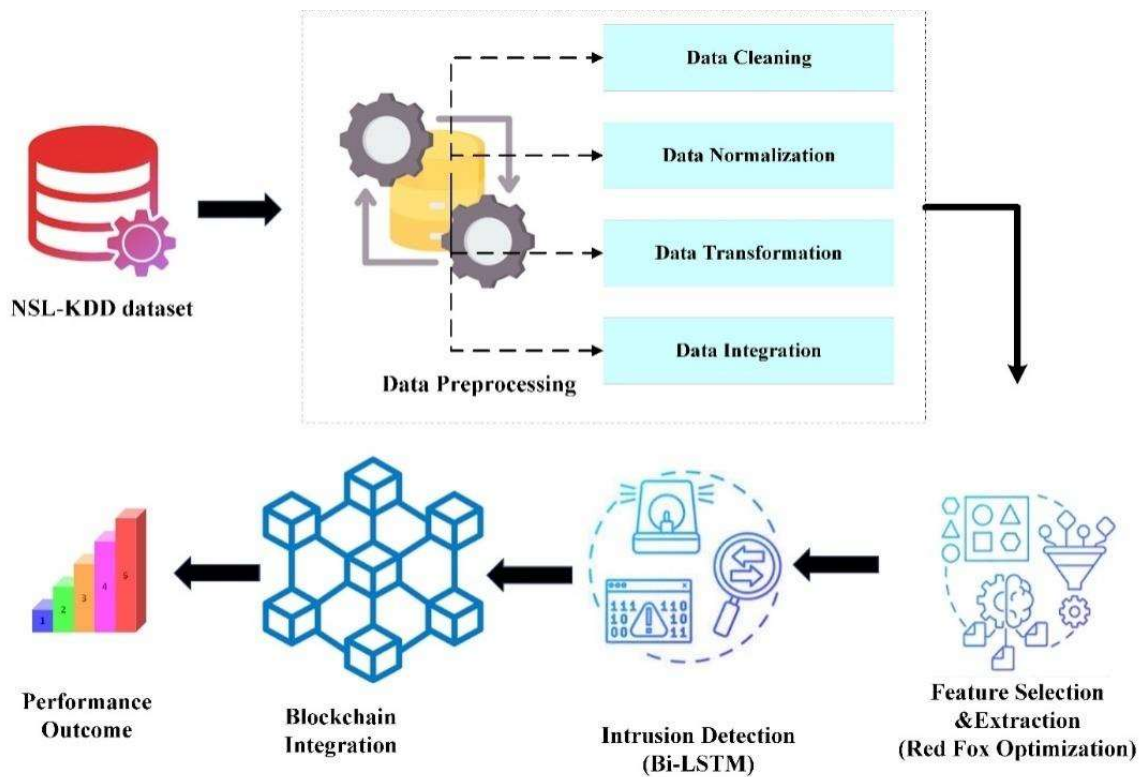


Fig. 1. Proposed integration of ML and blockchain for IoT intrusion detection.

### B. Data Preprocessing

Following data collection, the input data undergoes preprocessing to eliminate unwanted noise and address missing data. This involves four key preprocessing approaches:

- Data Cleaning
- Normalization
- Data Transformation
- Data Integration

### C. Data Cleaning

In order to improve the quality and dependability of datasets, data cleaning is an essential step in the data preparation pipeline. It involves locating and correcting different kinds of data abnormalities. These anomalies may include corrupted, incorrect, duplicate, or improperly formatted data entries. The primary goal of data cleaning is to ensure that datasets are standardized, accurate, and easily accessible for analysis and query purposes. During the data cleaning process, several tasks are performed to address different types of data issues. Firstly, corrupted or incorrect data entries are identified and either removed or corrected to restore data integrity. Duplicate entries, if present, are identified and eliminated to prevent redundancy and ensure that each observation is unique[18]. Additionally, managing missing values—which can occur for a number of reasons, including incomplete records or mistakes in data collection—is another aspect of data cleansing. When there are missing values in an observation, they can be imputed using statistical

techniques or data from other observations can be dropped. Additionally, data cleaning ensures that the dataset complies with the required format and schema by addressing structural flaws that could arise throughout the data transfer process. Thorough data cleaning improves the dataset's dependability and suitability for analysis, allowing analysts and researchers to derive precise conclusions and make defensible choices [19].

### D. Normalization

Normalization is a preprocessing step aimed at transforming data from its existing range to a new range. Given the presence of uncertain and incomplete data in the dataset, it becomes essential to address missing or irrelevant data to enhance data quality. The dataset can be integrated and normalized with success using the Min Max normalization approach. By making sure the dataset is scaled correctly, this method makes it possible to anticipate outcomes within the new range and allow for a greater difference in forecasting. Normalization reduces the influence of differences in dataset scales by scaling the dataset so that normalized values lie between 0 and 1. This allows for easier comparison of results from various datasets. This technique involves deducting the minimum value from the variable requiring normalization, resulting in a standardized dataset suitable for analysis and comparison. Min-max scaling, frequently referred to as feature scaling, converts the values of each feature to a range of 0 to 1 [20]. To compute the min-max scaling, use Eq. (1).

$$A_{scaled} = \frac{A - A_{min}}{A_{max} - A_{min}} \quad (1)$$

A is the starting value,  $A_{min}$  is the smallest value, and  $A_{max}$  is the largest value in the dataset. This method is helpful when the features are not evenly distributed and have a small range.

#### E. Data Transformation

Data transformation involves converting the original dataset into a specific format that facilitates faster and more efficient retrieval of strategic insights. Raw datasets can be challenging to comprehend and track, necessitating transformation into a more suitable form before extracting information. This transformation process is crucial for providing easily interpretable patterns, aligning with the strategic objectives of data conversion. Various techniques, such as smoothing, aggregation, and generalization, are employed in data transformation to streamline the dataset. Smoothing techniques are utilized to eliminate noise from the dataset, enhancing data clarity. Data aggregation gathers and presents data in a summarized format, aiding in easier analysis and interpretation. Additionally, data generalization involves converting lower-level or raw data into higher-level data through hierarchical concepts, further enhancing the dataset's organizational structure and usability [21].

#### F. Data Integration

Data integration is a preprocessing strategy that combines data from several sources into a single data repository to give rich views of the data. These sources could be flat files, databases, or several data cubes. Collaboration between users at all levels is facilitated by data integration, which combines received data with heterogeneous datasets to store consistent data that is client-accessible. A triplet defines the data integration mechanism, which is further explained in Eq. (2).

$$D_1 = \langle U, V, W \rangle \quad (2)$$

In this context,  $D_1$  represents the process of data integration, where U stands for the global schema, V denotes the schema of heterogeneous sources, and W refers to the mappings between queries of the source and global schema [22].

#### G. Feature Selection

In order to improve the effectiveness and productivity of the intrusion detection process, feature selection is an essential step in the preliminary processing phase of systems for detection. Its goal is to pick the most pertinent characteristics from the pre-processed data. Red Fox Optimisation (RFO) becomes apparent as a potent feature selection method in this scenario. To increase the intrusion detection system's overall performance, RFO works by optimising feature subsets. Finding a subset of characteristics that maximises the discrimination between normal and aberrant network behaviour is the main goal of feature selection using RFO. This will improve the system's capacity to detect intrusions effectively while reducing computing overhead. RFO does this by iteratively assessing and honing potential feature subsets according to pre-established optimisation standards, including performance metrics or classification accuracy. The intrusion detection system may efficiently prioritise and concentrate on the most useful aspects by using RFO for feature selection. This lowers the dimensionality of the data and boosts the

overall effectiveness of the detection process. Additionally, RFO has the flexibility and scalability to manage high-dimensional information that are frequently seen in Internet of Things networks [23].

After obtaining the balanced dataset from the previous stage, the optimal features for improving intrusion detection training speed and accuracy are selected using the DRF optimisation technique. Numerous meta-heuristic optimisation strategies are developed to improve network security in standard systems for detection of intrusions. Three newly created models used for network security are Spider Monkey Optimisation, Fruity Optimisation, and Greedy Swarm Optimisation. However, overfitting, which delayed processing, a slower rate of convergence, and complex computational procedures are the main causes of its issues. Generally speaking, some of the most current nature-inspired/bio-inspired optimisation approaches produced is the Dragon Fly Algorithm, Moth Flame Optimisation, and Ant Lion Optimisation, Harris Hawk optimisation (HHO), Flower Pollination Algorithm. These algorithms are commonly used to solve complex optimisation problems in a variety of security applications. The DRF is one of the newest optimisation algorithms and has several advantages over previous techniques. It has a low processing cost, less local optimum, rapid convergence, and guards against algorithm stacking during optimisation. Furthermore, the DRF35 is not specifically utilised in applications for IoT-IDS security. Therefore, the goal of the proposed study is to use this method to dataset feature optimisation based on the best optimum solution. Additionally, this optimisation procedure facilitates a simpler classification method with a higher assault detection rate [23].

The balanced IoT dataset's characteristics may be optimally tuned using this optimization approach. Foxes belong to many Canidae families and are tiny to medium-sized omnivore animals with pointed noses, long, thin legs, thicktails, and slender limbs. The foxes may also be distinguished from each other of their family and from large dogs. A novel meta-heuristic optimization system called the DRF takes its cues from the hunting habits of red foxes. When hunting, the red fox moves slowly towards its prey as it hides in the underbrush, and then it attacks the animal out of the blue. Like previous meta-heuristic models, this approach takes into account both the utilization and investigation of capabilities. This method creates random people for initializing parameters, as seen by the subsequent Eq. (3) and Eq. (4).

$$R = [r_0, r_1, \dots, r_{n-1}] \quad (3)$$

$$(R)^i = [(r_0)^i, (r_1)^i \dots (r_{n-1})^i] \quad (4)$$

where, "I" denotes how many populations are present in the search area. Ten, the global optimal function is used to find the best solution in the search space. Here, the structure that follows is used in conjunction with the Euclidean distance to get the best solution as presented in Eq. (5).

$$E(((R)^i)^k, (R_{best})^k) = \sqrt{(R^i)^k - (R_{best})^k} \quad (5)$$

In Eq. (5)  $k$  denotes the number of iterations. The term " $R_{best}^t$ " represents the best optimum, while " $E(.)$ " denotes the Euclidean distance. Accordingly, the optimal solution is employed to migrate all candidates, as illustrated in Eq. (6):

$$((R)^i)^k = ((R)^i)^{k-1} + g_{sign}((R_{best})^k - (R^i)^k) \quad (6)$$

As a scaling hyperparameter, " $g$ " denotes a random value selected at random from 0 to 1 for each iteration. For the whole population, this value is set just once every iteration. People evaluate the fitness values at their new places after moving to the optimal posture. People stay in their new roles if the fitness values are greater; if not, they return to their previous ones. This procedure is similar to how close relatives tell others where to hunt after an adventure and return home. They do what the explorers have instructed, going home "empty-handed" if they don't locate food, or continuing to search if there is a possibility. These processes, which take place during every DRF cycle, resemble suggested global inquiries. In addition, the applicants' move to new roles must present a feasible alternative; if not, their previous jobs will remain. The comparison of the red fox, advancing towards its prey and watches it, is appropriate here since it is similar to the DRF model in which a random number  $\omega$  between 0 and 1 is assumed explained in Eq. (7) and Eq. (8) [24].

$$\begin{cases} \text{Move Forward if, } \omega > \frac{3}{4} \\ \text{Stay Hidden if, } \omega > 3/4 \end{cases} \quad (7)$$

$$\omega = \begin{cases} h \times \frac{\sin(\delta_0)}{\delta_0} & \text{if } \delta_0 \neq 0 \\ \tau & \text{if } \delta_0 = 0 \end{cases} \quad (8)$$

Here, " $h$ " is a random number in the interval  $[0, 0.2]$ , and " $\delta_0$ " is another random number in the interval  $[0, 2\pi]$ , which indicates the fox viewing angle. Furthermore, " $\tau$ " represents a random number between 0 and 1. To model motions for the population of persons, the set of solutions for geographic coordinates is as follows. All things considered, the incorporation of RFO for picking features in intrusion detection systems improves computing efficiency and scalability while also strengthening the system's capacity to precisely detect and address security threats in Internet of Things networks. This method emphasises how crucial it is to use cutting-edge optimisation strategies in order to optimise feature subsets and improve intrusion detection technologies' overall effectiveness.

#### H. Intrusion Detection using Attention Bi-LSTM

The Attention-based BiLSTM model is used to identify intrusions in the NSL-KDD dataset. Using specialised memory units, LSTM—an improved version of the classic Recurrent Neural Networks (RNN)—captures long-term relationships in the MTS dataset efficiently [20]. The gradient vanishing problem is addressed by LSTM models, in contrast to conventional RNN techniques. Rather than depending just on the architecture of hidden units, they also incorporate memory cells that capture the long-term dependence of the signal. Four regulated gates make up the LSTM model: an output gate, a forget gate, input gate, in addition to a self-loop memory cell. These gates control how several memory neurons' data streams communicate with one another. The

forget gate in the LSTM model's hidden layer decides which data from the previous time frame to keep and which to discard. The input gate makes the decision to simultaneously inject data from the memory unit into the input signal or not. The output gate decides whether to change the state of the memory unit [24]. The following Eq. (9) through Eq. (14) are used to determine the neuron state, hidden layer results, and gate states, taking into account the input  $x_t$  from the NSL-KDD dataset and the dynamic output state  $h_t$ :

$$ip_t = \sigma(X_i u_t + Y_i h_{t-1} + a_i) \quad (9)$$

$$fg_t = \sigma(X_f u_t + Y_f h_{t-1} + a_f) \quad (10)$$

$$op_t = \sigma(X_o u_t + Y_o h_{t-1} + a_o) \quad (11)$$

$$c_t = fg_t \odot c_{t-1} + ip_t \odot \tilde{c}_t \quad (12)$$

The weight matrices that recur are indicated by as  $Y_i, Y_f, Y_o$ , while the representation of the weighted matrix for the forget, output, input, and memory cell gating by  $X_i, X_f, X_o$ , respectively. The biases for the gates are formulated as  $a_i, a_f, a_o$ . The candidate's cell state  $\tilde{c}_t$ , is utilized to update the original memory cell state,  $c_t$ . Step indicates the hidden layer's state  $h_{t-1}$  at any given moment, while  $ot$  indicates the output  $op_t$ . The symbol  $\odot$  denotes the element-wise multiplication operation. The hyperbolic tangent function is denoted as  $\tanh$ , and the logistic sigmoid activation function is represented by  $\sigma$ .

The standard LSTM model's limitation lies in its one-directional analysis of input signals during training, potentially leading to the inadvertent oversight of sequential information. In contrast, the BiLSTM was designed with a bidirectional structure, leveraging two LSTM layers operating in opposing directions to capture representation information both forwards and backwards. This bidirectional setup includes a hidden layer for reverse transmission (denoted as  $hb(t)$ ), incorporating future values, alongside a forward propagation hidden layer ( $hf(t)$ ) that retains data from previous sequence values. Ultimately, the BiLSTM model's final output is a fusion of both  $hf(t)$  and  $hb(t)$ , facilitating a more comprehensive understanding of time series data.

$$M_{fg}(t) = \varphi(Y_{fm} u_t + Y_{fmm} u_{f(t-1)} + a_{fa}) \quad (13)$$

$$M_a(t) = \varphi(Y_{am} u_t + Y_{amm} u_{a(t-1)} + a_a) \quad (14)$$

Besides these,  $a_{fa}$  and  $a_a$  also relate to two-way biased data. The weight matrix " $Y_{fm}$  and  $Y_{am}$ " represents the synaptic weights from the input value to the internal unit for both forward and backward directions. Similarly, the forward and backward feedback recurrent weights are denoted by  $Y_{fmm}$  and  $Y_{amm}$ .

The  $\tanh$  function serves as the activation function  $\psi$  for the hidden layers (HLs). It determines the output of the BiLSTM as  $b_t$ .

$$b_t = \sigma(W_{fmb} m_{f(t)} + W_{amb} m_{a(t)} + a_b) \quad (15)$$

The forward and backward weights of the resulting layers are represented by  $W_{fmb}$  and  $W_{amb}$ , respectively, in Eq. (15). Both a linear or sigmoidal function is provided as the



activation function of the resulting layer  $\sigma$ . Moreover,  $b$  denotes the bias in the output. The attention mechanism contributes to the learning process of the Attention BiLSTM model by assigning varying weights. The attention  $a_i$  for a hidden layer  $h_i$  is calculated using Eq. (16):

$$x_i = \tanh(Wh_i + a) \quad (16)$$

BiLSTM networks provide a powerful means to examine sequential data streams, enabling real-time detection of anomalous behavior and security threats in IoT networks. Leveraging BiLSTM architectures, these networks excel in capturing temporal dependencies and patterns present in IoT data, which are often characterized by their dynamic and time-varying nature. By effectively modelling the sequential nature of IoT data, BiLSTM networks can accurately identify deviations from normal behavior, facilitating prompt detection of intrusions and security breaches. To protect the integrity and confidentiality of IoT systems and devices, respond proactively to new threats, and strengthen the security posture of IoT networks, this capability is essential.

### I. Blockchain Integration

The integration of blockchain technology into intrusion detection systems involves several key steps to ensure the integrity and immutability of the data while facilitating secure communication among distributed IoT devices through smart contracts.

1) *Data logging*: In the process of data logging, intrusion detection data generated by IoT devices is systematically recorded onto the blockchain network. Each piece of data is meticulously timestamped and cryptographically secured, ensuring its integrity and safeguarding against any potential tampering attempts. By timestamping each entry, the blockchain network establishes a chronological order of events, enabling a comprehensive audit trail of intrusion activities. Additionally, the cryptographic security measures implemented within the blockchain network guarantee the immutability of the logged data, thereby providing a reliable and tamper-proof record of security events. This meticulous logging process enhances the trustworthiness and reliability of the intrusion detection system, enabling robust security monitoring in IoT networks [25].

2) *Blockchain node*: In the context of blockchain technology, blockchain nodes serve as essential components responsible for validating and recording logged intrusion detection data. These nodes are distributed across the blockchain network, ensuring decentralization and resilience against single points of failure. Each node maintains a copy of the decentralized ledger, which contains a complete record of all transactions, including the logged intrusion detection data. When new data is logged onto the blockchain, it undergoes validation by multiple nodes within the network to ensure its authenticity and integrity. This validation process involves verifying the cryptographic signatures associated with the data and confirming its adherence to the consensus rules established by the network protocol. Once validated, the intrusion detection data is appended to the blockchain ledger,

becoming a permanent and immutable part of the distributed database. By distributing the responsibility for data validation and storage among multiple nodes, blockchain networks achieve redundancy and fault tolerance, enhancing the reliability and resilience of the overall system. Furthermore, as blockchain nodes are decentralised, no one organisation can exert control over the system as a whole, fostering openness, confidence, and security in the logging and archiving of intrusion detection data.

3) *Proof of work*: The consensus mechanism of the blockchain is essential to guaranteeing that all dispersed nodes agree on the veracity of logged data. To reach this consensus among network users, consensus techniques like Proof of Work (PoW) are used. Proof-of-work (PoW) consensus is a competitive mechanism in which nodes solve challenging mathematical problems to validate transactions and append new blocks to the blockchain. This is a resource-intensive procedure that uses a lot of energy and processing power. Nonetheless, other nodes in the network confirm the answer after a node completes the puzzle and suggests a new block. The block is appended to the blockchain if the answer satisfies the consensus requirements. By using this decentralised method, blockchain networks maintain the integrity and durability of the blockchain ledger by facilitating consensus across dispersed nodes about the veracity of recorded data. Additionally, consensus mechanisms like PoW contribute to the security of the blockchain network by mitigating the risk of malicious actors attempting to manipulate or alter the logged data. Overall, the consensus mechanism serves as a fundamental building block of blockchain technology, enabling decentralized trust and coordination among network participants [26].

A key element of blockchain networks is the proof-of-work (PoW) consensus mechanism, which guarantees dispersed nodes' agreement on the legitimacy of transactions and the appending of new blocks to the blockchain. PoW comprises the following crucial steps:

- **Transaction Propagation**: Transactions are broadcasted to all nodes in the blockchain network. Each transaction contains details such as sender, recipient, amount, and cryptographic signatures.
- **Block Creation**: Transactions are grouped together into blocks, forming a candidate block for addition to the blockchain. Miners, who are nodes responsible for creating new blocks, select transactions and assemble them into a block structure.
- **Mining Competition**: Miners compete with each other to solve the Proof of Work puzzle. They utilize computational power to generate hash values by iteratively modifying a nonce (a random number) in the block header until the desired hash value is found. This process is computationally intensive and requires significant computational resources.

- **Verification:** A miner broadcasts the candidate block and the solution to the network as soon as they discover a workable solution to the problem. The legitimacy of the answer and the transactions included in the block are then confirmed by further nodes inside the network.
- **Consensus:** If the majority of nodes in the network agree that the proposed solution is sound and the block conforms to the consensus requirements, the block is accepted and posted to the blockchain. It is ensured that all distributed nodes concur on the validity of the transactions and the addition of new blocks to the blockchain by going through this process.
- **Reward:** A fixed quantity of bitcoin plus any transaction fees included in the block are awarded to the miner who effectively mines a new block. This encourages miners to use up processing power and take part in the consensus-building process on the network.

In general, the Proof of Work technique reduces the possibility of malevolent actors attempting to influence the blockchain by demanding computational resources to verify transactions and generate new blocks, hence ensuring the security and integrity of blockchain networks.

1) *Smart contract:* Smart contracts serve as the backbone of automation and governance within IoT networks by providing a decentralized, programmable framework for enforcing rules and conditions. These contracts, encoded with predefined logic, are deployed on the blockchain, ensuring immutability and tamper-proof execution. Within the context of IoT, smart contracts automate interactions between devices, enabling seamless communication and coordination without the need for intermediaries. By executing automatically when specific conditions are met, such as sensor readings or trigger events, smart contracts streamline processes and mitigate the risk of human error. Moreover, the decentralised structure of these systems gets rid of single points of failure and minimises dependence on centralised authority, hence improving security and resilience. Additionally, conditional execution of operations is made possible by smart contracts, which let gadgets react quickly to shifting conditions. This feature improves IoT network responsiveness and operational efficiency. Furthermore, network participants' confidence and responsibility are bolstered by the openness and auditability provided by smart contracts. Overall, smart contracts play a critical role in driving efficiency, security, and transparency in IoT ecosystems, laying the foundation for scalable and resilient decentralized applications [27].

2) *Secure communication:* In the ecosystem of IoT networks, secure communication is facilitated through the interaction between IoT devices and the blockchain network via smart contracts. These contracts act as intermediaries, enforcing cryptographic protocols and access controls to ensure that communication remains secure. By leveraging cryptographic techniques such as encryption and digital signatures, smart contracts authenticate and authorize devices,

mitigating the risk of unauthorized access or tampering. Through predefined rules and conditions encoded within the smart contracts, only authorized devices are granted permission to access and modify data stored on the blockchain. This robust enforcement of security measures enhances the integrity and confidentiality of communication within IoT networks, safeguarding sensitive information and preventing unauthorized manipulation of data. Overall, the utilization of smart contracts enables secure and trustworthy communication channels, fostering confidence in the exchange of data and transactions within IoT ecosystems.

## V. RESULT AND DISCUSSION

The proposed framework undergoes rigorous evaluation using benchmark datasets, including NSL-KDD and BoT-IoT, to comprehensively assess its performance in detecting various types of intrusions within IoT networks. By leveraging these datasets, which contain diverse and realistic intrusion scenarios, the framework's efficacy in identifying and mitigating security threats is thoroughly scrutinized. Performance metrics are used to assess how well the framework differentiates between malicious activity and typical network behavior. These measures include detection accuracy, false positive rate, and computing efficiency. Furthermore, the assessment procedure entails contrasting the outcomes of the framework with those of current intrusion detection systems in order to measure its effectiveness in relation to predetermined benchmarks. The suggested framework's potential to strengthen the security posture of IoT networks is carefully investigated through this methodical study utilizing typical datasets, offering insights into its advantages and shortcomings.

### A. Performance Metrics

Performance metrics refer to the numerical values that are utilized to assess how well an intrusion detection system detects and neutralizes security threats on a network. Commonly used metrics include the following ones:

1) *Accuracy:* The percentage of accurately identified occurrences—both true positives and true negatives—out of all the instances that were examined is known as accuracy. It offers a general indicator of how effectively the intrusion detection system classifies events as either intrusions or routine activity.

2) *Precision:* Positive predictive value, or precision, is a metric that expresses the percentage of accurately detected positive cases (true positives) across every case categorized as positive (false positives and true positives). It shows how well the system can detect intrusions without mistakenly labelling routine operations as such.

3) *Recall:* Recall, also known as sensitivity or true positive rate, is the proportion of correctly identified positive cases relative to all real positive occurrences in the dataset. It assesses the system's ability to identify every incursion, lowering the likelihood that any malicious activity would go undetected.

4) *F1-score*: The F1-score, which achieves equilibrium between recall and accuracy, is derived from the harmonic mean of these two metrics. Recall and accuracy are combined into one figure, which accounts for both false positives and false negatives.

TABLE I. PERFORMANCE METRICS

Metrics	Efficiency
Accuracy	98.9
Precision	94
Recall	95
F1-Score	95

As shown in Table I and Fig. 2, the suggested intrusion detection approach exhibits excellent efficiency with an accuracy of 98.9%, demonstrating its capacity to accurately categorise cases as either intrusions or routine operations. Furthermore, the approach displays a 94% accuracy rate, which indicates the percentage of accurately detected incursions among all cases that are categorised as positive, hence reducing false positives. With a recall rate of 95%, which indicates that the system can detect all incursions, there is little chance of a missed detection. Furthermore, a balanced performance in terms of both accuracy and recall is shown by the F1-score, which harmonises the two metrics, which is recorded at 95%. All of these measures show how successful and dependable the suggested intrusion detection technique is at identifying and reducing security risks in the network infrastructure.



Fig. 2. Performance efficiency.

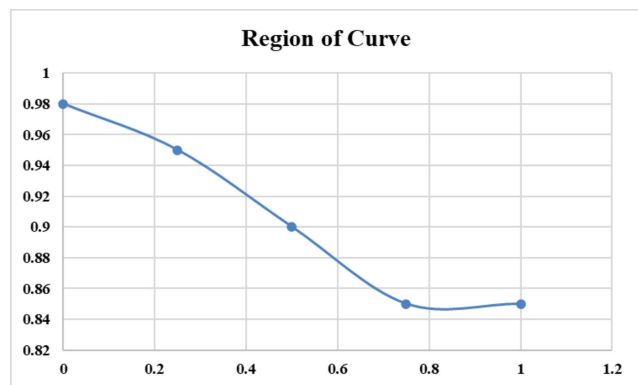


Fig. 3. Receiver operating characteristic curve.

As the threshold rises from 0 to 1, the true positive rate (TPR) progressively falls from 0.98 to 0.85, suggesting a decline in the percentage of true positive cases that are correctly categorised, as seen in Fig. 3. The TPR stays comparatively high at 0.95 at a threshold of 0.25, indicating that true positive cases can be effectively detected even with somewhat loosened thresholds.

TABLE II. SORTING RESULT OF NSL-KDD

Methods	AUC	Error Rate
Gradient Boosting Classifier	47.64	0.4905
Deep Learning	77.88	0.2256
Proposed Method	98.9	0.0025

The NSL-KDD dataset's categorization outcomes using different techniques are shown in Table II. With an error rate of 0.4905 and an AUC of 47.64%, the Gradient Boosting Classifier performs relatively poorly. By comparison, the Deep Learning approach shows noticeably higher performance, with an error rate of 0.2256 and an AUC of 77.88%.

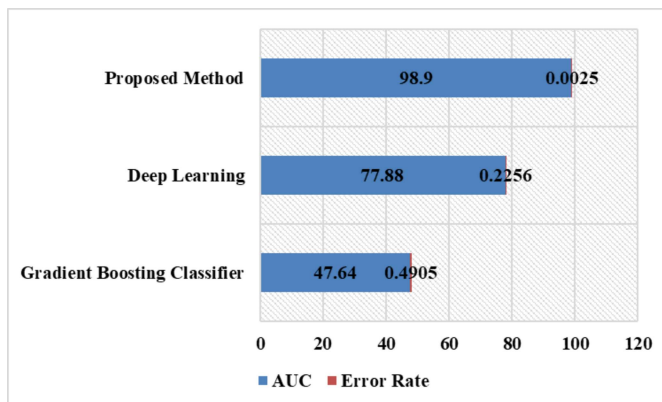


Fig. 4. Classification result of NSL-KDD.

The results presented in Fig. 4 demonstrate that the suggested approach outperforms the two other options, with an exceptional AUC of 98.9% and a remarkably low error rate of 0.0025. These outcomes highlight how well the suggested strategy performs in comparison to other methods when it comes to correctly identifying instances in the NSL-KDD dataset.

TABLE III. RECOGNITION OUTCOMES OF ATTENTION BASED BiLSTM APPROACH ON NSL-KDD DATASET

Data	Class	Accuracy	Precision	Recall	F1-Score
Training	Normal	98.4	96.3	97.3	96.3
	Attack	97.4	97.4	98.4	95.5
	Average	97.7	97.7	97.7	97.7
Testing	Normal	98.9	97.5	96.4	95.8
	Attack	97.3	98.3	97.3	98.3
	Average	98.9	98.9	98.9	98.9

The NSL-KDD dataset's recognition results from the Attention-based BiLSTM technique are shown in Table III and Fig. 5.

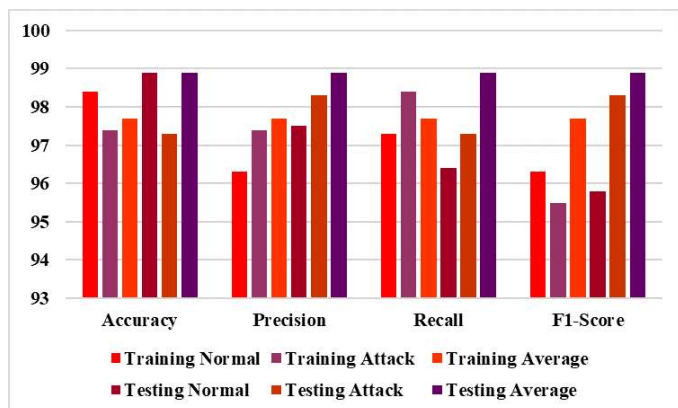


Fig. 5. Recognition outcomes of attention based BiLSTM approach on NSL-KDD dataset.

It includes accuracy, precision, recall, and F1-Score, split into normal and attack classes, for both the training and testing datasets. In the training dataset, the method achieves 98.4% accuracy for normal cases and 97.4% accuracy for attack instances. Respectively, the corresponding accuracy, recall, and F1-Score values are 95.5% and 96.3%, 97.3% and 98.4%, and 96.3% and 97.4%. Comparable outcomes are seen in the testing dataset, where the technique achieves 97.3% accuracy for attack instances and 98.9% accuracy for normal cases. The corresponding F1-Score, recall, and accuracy scores are 97.5%, 98.3%, and 95.8%, respectively. The average results for each class are also provided for the training and testing datasets.

### B. Discussion

The research studies under discussion offer novel strategies for resolving the security issues that arise in Internet of Things networks. These specifically concentrate on intrusion detection systems (IDS) and make utilisation of blockchain and machine learning technology. The study by Liang et al. [12] suggests a hybrid intrusion detection system that makes use of multi-agent systems, blockchain technology, and deep learning techniques. The system is divided into distinct modules for data collecting, management, analysis, and response with the goal of improving detection accuracy, particularly at the transport layer of Internet of Things networks. Scalability and optimisation continue to be major obstacles to practical implementation, notwithstanding encouraging findings. A collaborative intrusion detection architecture including blockchain technology for safe sharing of threat intelligence across cloud and Internet of Things networks is presented by Alkadi et al. [13]. While consensus processes and deep blockchain technology are adept at detecting intrusions and reducing security threats, their scale presents serious problems for efficiency and real-time response. A blockchain-based collaborative signature-based IDS called CBSigIDS is proposed by Li et al. with the goal of creating a trustworthy signature database in dispersed IoT systems. Although it provides a safe way to validate signatures, blockchain overhead scalability issues require

further work before a viable implementation can be made. Kumar et al. [14] offers a distributed intrusion detection system (IDS) that uses blockchain technology and fog computing to identify DDoS assaults directed at IoT mining pools. They assess the system's effectiveness in identifying IoT network assaults using machine learning algorithms trained on scattered fog nodes. But there are still issues with realistic implementation and optimisation needed for efficiency and scalability. Although these studies show how blockchain and machine learning technologies could potentially use to improve IoT network security, scalability, optimisation, and practical deployment issues must be resolved before their full promise could be realised in practical settings.

The study presents a complete framework for reliable and scalable intrusion detection in IoT networks by integrating machine learning techniques with blockchain technology. The solution addresses the challenges posed by the dynamic and heterogeneous nature of IoT environments by employing Red Fox Optimization for feature selection and Attention-based BiLSTM for anomaly identification. The adoption of blockchain technology improves security by ensuring the validity and inviolability of intrusion record detection. The study advances the area by providing an all-encompassing method of intrusion detection that takes security and efficiency into account. Real-time identification of abnormalities and malicious activity in IoT traffic is made possible by the use of sophisticated machine learning algorithms, and scalability is improved by optimization approaches that assist decrease the dimensionality of the input data. Furthermore, the system gains an additional degree of protection through the integration of the technology known as blockchain, which offers tamper-resistant recordings of detected intrusions. The usefulness of the suggested architecture is demonstrated by experimental findings, which on real-world IoT datasets yield a high detection accuracy of about 98.9%. These findings highlight how important the study is to improving IoT security state-of-the-art. The report does, however, admit several limitations, including the need for more assessment in various IoT scenarios and the computational cost related to blockchain integration. Prospective study avenues encompass investigating alternative machine learning algorithms and optimization methods, tackling scalability issues, and refining blockchain-associated procedures. Overall, the research offers a viable strategy for improving intrusion detection in Internet of Things networks, opening the door to more robust and safe linked settings.

## VI. CONCLUSION

The suggested system, which makes use of blockchain and machine learning, offers a viable solution to the problems associated with intrusion detection in Internet of Things networks. The accuracy and scalability of the intrusion detection system are improved by integrating Red Fox Optimization for feature selection and Attention-based BiLSTM for anomaly detection. Moreover, the incorporation of blockchain technology ensures the integrity and immutability of intrusion detection logs, thereby enhancing security. On real-world IoT data sets, experimental findings show the usefulness of the technique with a high detection



accuracy of about 98.9%. However, it is important to acknowledge some limitations and areas for future work. Firstly, while the proposed framework shows promising results, further research is needed to evaluate its performance in diverse IoT environments and under various attack scenarios. Additionally, the scalability of the system needs to be investigated to handle large-scale IoT networks efficiently. Furthermore, the computational overhead associated with blockchain integration may pose challenges in resource-constrained IoT devices, requiring optimization strategies. Moreover, continuous advancements in intrusion techniques necessitate ongoing updates and improvements to the detection algorithms and feature selection methods. Future studies may look at applying more machine learning algorithms and optimization techniques to enhance the robustness and efficiency of intrusion detection systems in Internet of Things networks. All things considered, this work establishes the groundwork for next investigations that seek to create IoT ecosystems that are more robust and safer.

#### REFERENCES

- [1] P. Raj and A. C. Raman, *The Internet of Things: Enabling technologies, platforms, and use cases*. Auerbach Publications, 2017.
- [2] V. E. Balas and S. Pal, *Healthcare Paradigms in the Internet of Things Ecosystem*. Academic Press, 2020.
- [3] Y. Liao, C. Thompson, S. Peterson, J. Mandrola, and M. S. Beg, "The future of wearable technologies and remote monitoring in health care," *Am. Soc. Clin. Oncol. Educ. Book*, vol. 39, pp. 115–121, 2019.
- [4] A. Karale, "The challenges of IoT addressing security, ethics, privacy, and laws," *Internet Things*, vol. 15, p. 100420, 2021.
- [5] A. Qasem, P. Shirani, M. Debbabi, L. Wang, B. Lebel, and B. L. Agba, "Automatic vulnerability detection in embedded devices and firmware: Survey and layered taxonomies," *ACM Comput. Surv. CSUR*, vol. 54, no. 2, pp. 1–42, 2021.
- [6] R. Krishnamurthi, A. Kumar, D. Gopinathan, A. Nayyar, and B. Qureshi, "An overview of IoT sensor data processing, fusion, and analysis techniques," *Sensors*, vol. 20, no. 21, p. 6076, 2020.
- [7] A. Riah, S. Daniel, E. Frank, and K. Seriffdeen, "The role of technology in shaping user behavior and preventing phishing attacks," 2024.
- [8] T. M. Alshammari and F. M. Alserhani, "Scalable and Robust Intrusion Detection System to Secure the IoT Environments using Software Defined Networks (SDN) Enabled Architecture," *Int J Comput Netw. Appl.*, vol. 9, no. 6, pp. 678–688, 2022.
- [9] M. Javed, N. Tariq, M. Ashraf, F. A. Khan, M. Asim, and M. Imran, "Securing Smart Healthcare Cyber-Physical Systems against Blackhole and Greyhole Attacks Using a Blockchain-Enabled Gini Index Framework," *Sensors*, vol. 23, no. 23, p. 9372, 2023.
- [10] A. Laszka, A. Dubey, M. Walker, and D. Schmidt, "Providing privacy, safety, and security in IoT-based transactive energy systems using distributed ledgers," in *Proceedings of the Seventh International Conference on the Internet of Things*, 2017, pp. 1–8.
- [11] A. K. Al Hwaitat et al., "A New Blockchain-Based Authentication Framework for Secure IoT Networks," *Electronics*, vol. 12, no. 17, p. 3618, Aug. 2023, doi: 10.3390/electronics12173618.
- [12] C. Liang et al., "Intrusion Detection System for the Internet of Things Based on Blockchain and Multi-Agent Systems," *Electronics*, vol. 9, no. 7, p. 1120, Jul. 2020, doi: 10.3390/electronics9071120.
- [13] O. Alkadi, N. Moustafa, B. Turnbull, and K.-K. R. Choo, "A deep blockchain framework-enabled collaborative intrusion detection for protecting IoT and cloud networks," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9463–9472, 2020.
- [14] R. Kumar, P. Kumar, R. Tripathi, G. P. Gupta, S. Garg, and M. M. Hassan, "A distributed intrusion detection system to detect DDoS attacks in blockchain-enabled IoT network," *J. Parallel Distrib. Comput.*, vol. 164, pp. 55–68, Jun. 2022, doi: 10.1016/j.jpdc.2022.01.030.
- [15] H. Vargas, C. Lozano-Garzon, G. A. Montoya, and Y. Donoso, "Detection of Security Attacks in Industrial IoT Networks: A Blockchain and Machine Learning Approach," *Electronics*, vol. 10, no. 21, p. 2662, Oct. 2021, doi: 10.3390/electronics10212662.
- [16] R. H. Hylock and X. Zeng, "A Blockchain Framework for Patient-Centered Health Records and Exchange (HealthChain): Evaluation and Proof-of-Concept Study," *J. Med. Internet Res.*, vol. 21, no. 8, p. e13592, Aug. 2019, doi: 10.2196/13592.
- [17] "NSL-KDD." Accessed: Mar. 21, 2024. [Online]. Available: <https://www.kaggle.com/datasets/hassan06/nslkdd>.
- [18] H. Moudoud, S. Cherkaoui, and L. Khoukhi, "An IoT blockchain architecture using oracles and smart contracts: the use-case of a food supply chain," in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, IEEE, 2019, pp. 1–6.
- [19] P. Bari and P. Karande, "Application of PROMETHEE-GAIA method to priority sequencing rules in a dynamic job shop for single machine," *Mater. Today Proc.*, vol. 46, pp. 7258–7264, 2021, doi: 10.1016/j.matpr.2020.12.854.
- [20] P. Yazdani and S. Sharifian, "E2LG: a multiscale ensemble of LSTM/GAN deep learning architecture for multistep-ahead cloud workload prediction," *J. Supercomput.*, vol. 77, pp. 11052–11082, 2021.
- [21] F. Karim, S. Majumdar, and H. Darabi, "Insights Into LSTM Fully Convolutional Networks for Time Series Classification," *IEEE Access*, vol. 7, pp. 67718–67725, 2019, doi: 10.1109/ACCESS.2019.2916828.
- [22] Z. Ahamed, M. Khemakhem, F. Eassa, F. Alsolami, and A. S. A.-M. Al-Ghamdi, "Technical Study of Deep Learning in Cloud Computing for Accurate Workload Prediction," *Electronics*, vol. 12, no. 3, p. 650, 2023.
- [23] E. S. P. Krishna and A. Thangavelu, "Attack detection in IoT devices using hybrid metaheuristic lion optimization algorithm and firefly optimization algorithm," *Int. J. Syst. Assur. Eng. Manag.*, May 2021, doi: 10.1007/s13198-021-01150-7.
- [24] R. AlGhamdi, "Design of Network Intrusion Detection System Using Lion Optimization-Based Feature Selection with Deep Learning Model," *Mathematics*, vol. 11, no. 22, p. 4607, Nov. 2023, doi: 10.3390/math11224607.
- [25] S. Rathore, J. H. Park, and H. Chang, "Deep Learning and Blockchain-Empowered Security Framework for Intelligent 5G-Enabled IoT," *IEEE Access*, vol. 9, pp. 90075–90083, 2021, doi: 10.1109/ACCESS.2021.3077069.
- [26] A. H. Sodhro, S. Pirbhulal, M. Muzammal, and L. Zongwei, "Towards blockchain-enabled security technique for industrial internet of things based decentralized applications," *J. Grid Comput.*, vol. 18, pp. 615–628, 2020.
- [27] B. Yin, Y. Wu, T. Hu, J. Dong, and Z. Jiang, "An efficient collaboration and incentive mechanism for Internet of Vehicles (IoV) with secured information exchange based on blockchains," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1582–1593, 2019.

## RESEARCH ARTICLE

# An opportunistic energy-efficient dynamic self-configuration clustering algorithm in WSN-based IoT networks

Sridevi Tumula<sup>1</sup> | Y. Ramadevi<sup>2</sup> | E. Padmalatha<sup>1</sup> | G. Kiran Kumar<sup>1</sup> |  
 M. Venu Gopalachari<sup>3</sup> | Laith Abualigah<sup>4,5,6,7,8,9</sup> | Premkumar Chithaluru<sup>1</sup>  |  
 Manoj Kumar<sup>10,11</sup> 

<sup>1</sup>Department of Computer Science and Engineering, Chaitanya Bharathi Institute of Technology, Hyderabad, India

<sup>2</sup>Department of AIML, Chaitanya Bharathi Institute of Technology, Hyderabad, India

<sup>3</sup>Department of Information Technology, Chaitanya Bharathi Institute of Technology, Hyderabad, India

<sup>4</sup>Computer Science Department, Al al-Bayt University, Mafraq, Jordan

<sup>5</sup>Department of Electrical and Computer Engineering, Lebanese American University, Byblos, Lebanon

<sup>6</sup>Hourani Center for Applied Scientific Research, Al-Ahliyya Amman University, Amman, Jordan

<sup>7</sup>Applied Science Research Center, Applied Science Private University, Amman, Jordan

<sup>8</sup>School of Computer Sciences, Universiti Sains Malaysia, Pulau Pinang, Malaysia

<sup>9</sup>School of Engineering and Technology, Sunway University Malaysia, Petaling Jaya, Malaysia

<sup>10</sup>School of Computer Science, FEIS, University of Wollongong in Dubai, Dubai, United Arab Emirates

<sup>11</sup>MEU Research Unit, Middle East University, Amman, Jordan

## Correspondence

Manoj Kumar, School of Computer Science, FEIS, University of Wollongong in Dubai, Dubai Knowledge Park, Dubai, United Arab Emirates.

Email: [wss.manojkumar@gmail.com](mailto:wss.manojkumar@gmail.com)

## Summary

The demand for the Internet of Things (IoT) has significantly increased in the current scenario; specific sectors that require IoT include industrial automation, home control, health care applications, military and surveillance applications, habitat monitoring, and nanoscopic sensor applications. The use of optimal wireless sensor networks (WSNs) in transmission techniques has resulted in their involvement. A WSN is made up of thousands of randomly distributed sensor nodes that sense and transmit environmental data such as temperature, pressure, humidity, light, and sound. One of the most important requirements when using these sensor nodes is energy. As a result, it has become a major area of research in recent years; additionally, several design techniques and protocols have been presented in the last decade, particularly for IoT-based applications. As a result, the systemization of an energy-optimized WSN in dynamic functional conditions with automatic self-configuration of sensor nodes is a critical goal. This paper proposed an opportunistic energy-efficient dynamic self-configuration routing (OEDSR) algorithm for IoT-based applications. Initially, the optimal route to the base

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2023 The Authors. *International Journal of Communication Systems* published by John Wiley & Sons Ltd.

station (BS) is calculated by using the residual energy and mobility factor of the sensor nodes obtained through a routing tree model based on graph theory. To reduce the number of connections, an optimal path is determined based on dynamic cluster generation through a hierarchical tree architecture. Finally, the network-related parameters, such as throughput, delay and packet delivery ratio (PDR), are compared with the peer existing routing protocols to depict the efficiency of the OEDSR protocol.

#### KEYWORDS

energy-efficient, IoT, mobility, PDR, self-configuration, WSN

## 1 | INTRODUCTION

A typical WSN is likely to involve elemental blocks that are required to observe, process, and communicate with neighboring nodes. These blocks allow users to sense, act on, and transmit information based on the needs of the identified circumstances. The user could be someone involved in business, medicine, civil, or government operations. A framework, biological system, or physical environment may be used to describe the conditions.<sup>1</sup> Surveillance, telemedicine, data collection, and forest fire monitoring are just a few of the applications of WSN-IoT applications. The WSN's purpose is not only to observe and transmit information but also to control information remotely using actuators.<sup>2</sup>

The WSN components are classified into four major categories: (1) wireless sensor nodes distributed at random, (2) gateway connection, (3) cluster heads (CHs), and (4) BS. The CH is a cluster node responsible for collecting data from all of the sensors in its cluster and relaying it to the BS.<sup>2</sup> Because the WSN collects a massive amount of data, data processing and analysis are considered critical events.<sup>3</sup> The computational and infrastructural designs of WSN-based IoT vary depending on the environmental specifications.<sup>4</sup> WSNs are typically expected to operate in a restricted environment. As a result, a difficult wireless structure was formed, and constraints were examined.<sup>5</sup> However, the WSN has dealt with well-known routing schemes; additionally, specific ad hoc protocols are not permitted within the WSN.

### 1.1 | Communications in IoT

Intra-cluster communication occurs within the cluster, and inter-cluster communication occurs with neighboring cluster sensor nodes.<sup>6</sup> Sensing, radio channel observation, and operations with computational constraints are some of the high-power-consuming operations of WSN nodes. Sensor nodes emit power not only during data packet reception and transmission but also when they simply listen to the network channel for packet routing information.

The sensor node's responsibility is to collect and transmit information to the cluster's CH. Furthermore, during the clustering of the sensor nodes, the physical location of the sensor nodes must be communicated to all sensor nodes in the nearby clusters in order to inform the clusters to combine the sensor nodes during the clustering process.<sup>7</sup> The BS node is in charge of selecting sensor nodes within clusters to sense information from the environment and transmit it to the corresponding CH.

Certain factors may affect the communication path during the wireless channel communication, resulting in congestion or bottleneck. The bottleneck problem arises as a result of inefficient clustering mechanisms.<sup>8</sup> Congestion or bottlenecking has the effect of increasing delay and decreasing PDR and throughput. A collision-free traffic-aware routing scheme must be implemented to achieve a reliable communication channel. While considering the data, the system performs a data segregation process due to the repeated transmission of redundant data over the network. The BS node<sup>9</sup> deletes the repeated data during this process. However, it raises the computational complexity of the BS node. To avoid this, all sensor nodes that are extremely close together are not allowed to sense and transmit data.

## 1.2 | Clustering in IoT

The BS node is responsible for informing the user about the value of the collected data; additionally, transmitting the aggregated data to the destination sensor node reduces computational complexity and system load.<sup>10</sup> Sensor node clustering is used to reduce the complexity of data aggregation. To manipulate the WSN stream, several clustering mechanisms are available. According to the simulation results of one study, a network with efficient clustering can provide twice the lifetime of non-optimized WSNs. Clustering also improves network scalability. Figure 1 illustrates the clustering in IoT.

The CHs are chosen based on the node parameters. The parameters determining the CHs are node mobility, node location, and the node's average residual energy (ARE).<sup>11</sup> For the duration of the sequence of operations, the selected CH remains the head. Clustering methods are classified into two types: those with a fixed cluster size and those with a variable cluster size. The static cluster size is used in stable WSNs, while the dynamic cluster size is used in mobile WSNs.<sup>12</sup> The BS node also provides the cluster's minimum threshold size.<sup>13</sup> The network topology varies depending on the mobility of the nodes; thus, the cluster size varies from one set of operations to another.

Because these sensor nodes are processed to combine as a group of multiple clusters, enlarging the network in the event of an association with any other conjugate network<sup>14</sup> is much easier. Furthermore, because the infrastructure required for the wireless network is significantly less expensive and more affordable, network augmentation is simple. Figure 2 represents inter and intra-cluster communications.

Implementing a clustering algorithm is a difficult task. Following the implementation of the clustering algorithm, the network may face the following design and processing challenges that affect clustering in a WSN.

1. Data collection: Normally, the physical design of sensor nodes is limited to ensure compactness. Because they are mostly used in unattended areas, they must be cost-effective. As a result, one of the major factors limiting clustering performance in a WSN is the constraint of data accumulation. For example, when certain network-related issues

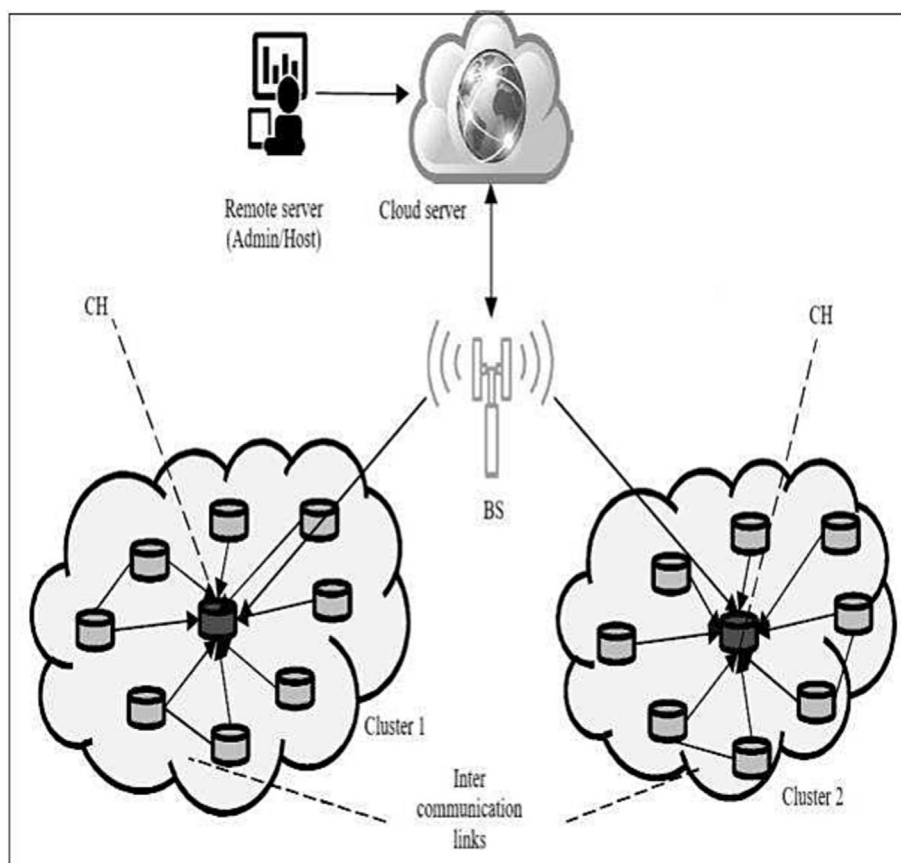


FIGURE 1 Clustering in IoT.



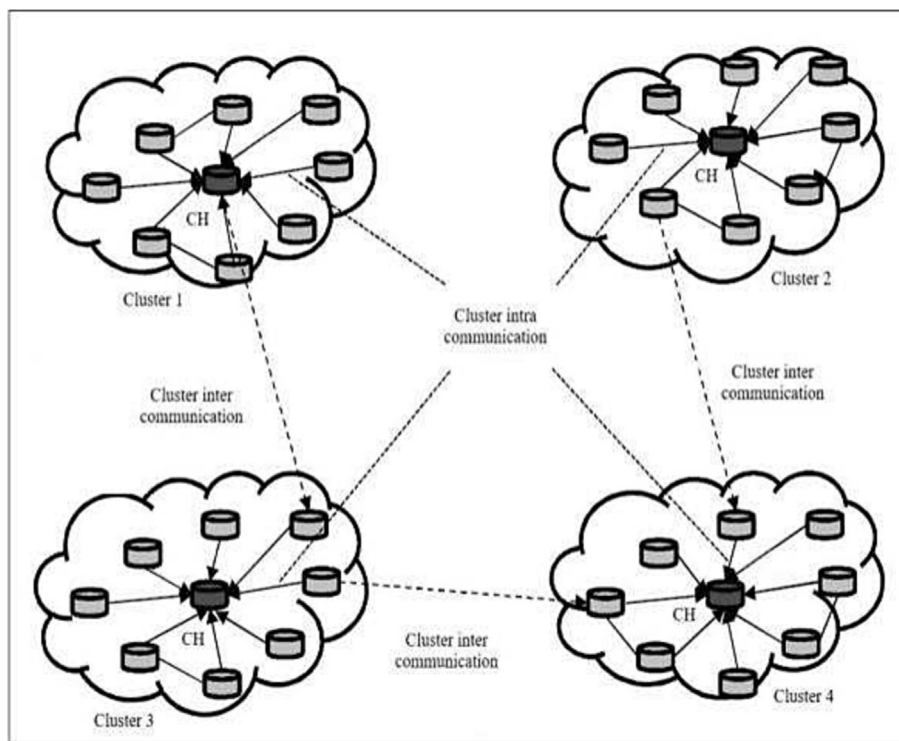


FIGURE 2 Inter and intra cluster communication.

occur, a dependable WSN may be required to store specific packets. In this case, the sensor node can handle small-sized data packets; however, if the data size exceeds the limit, the packet is dropped, and the clustering algorithm is deemed unfit for implementation.

2. Data distribution: The data to be distributed must traverse the network while considering network coverage. Every sensor node in the cluster must respond appropriately in this case. When there is a discrepancy in the clustering technique, certain sensor nodes remain uncombined. As a result, if the information reaches the uncombined sensor nodes, these nodes will attempt to communicate directly with the BS.<sup>15</sup> If the BS node is far from the transmitting node, the sensor node's energy consumption will be high. Furthermore, the likelihood of the data reaching the BS is low.
3. Security concerns: Because WSN-IoT nodes are vulnerable to a wide range of security threats, a clustering algorithm must be built to withstand malicious attacks. Attacks can occur anywhere within the cluster.<sup>16</sup> During the clustering process, for example, a malicious node may attempt to combine with a cluster. If the algorithm is security-specific, the malicious node will be identified and removed from the WSN.
4. Reserved energy: Transmitting a single data packet through a non-clustered WSN allows the data to travel at a high communication cost; additionally, data loss or leakage may occur in this type of communication.<sup>17</sup> Clustering would alleviate the problem because the CH would be the next-hop destination for every sensor node in the cluster; thus, the data packet could identify the ideal destination for the next hop, preventing the looping problem. As a result, balanced energy utilization is impossible without a proper clustering process.
5. Network lifetime: If the WSN is not clustered, the sensor node's energy is depleted. This does not imply that data transmission is impossible in the absence of clustering.<sup>18</sup> However, without clustering, sensor nodes lose energy quickly and transmission times are extremely long. Because the WSN's lifetime is directly proportional to the energy consumption of the sensor nodes, the clustering process is critical for optimizing network lifetime.

### 1.3 | Security challenge in IoT

Because of its ease of construction and low technical requirements, the WSN is easily vulnerable to security threats. As a result, security is a major concern in the case of highly confidential applications, such as military and medical

diagnostics, such as telemedicine. Because of the characteristics of the WSN, traditional security techniques are unable to overcome security threats.<sup>19</sup> There are several ways for an attacker to gain access to the WSN.

The cluster's CHs send and receive data to the BS node. Rather than attacking the sensor node, the attacker can inject malicious data into it. Furthermore, the BS node cannot determine whether or not the information received is reliable. Certain silent attacks may occur in the WSN; these attacks are extremely harmful because they generate bogus data<sup>20</sup> into the network routing stream. This unwanted redundant data causes cluster congestion or bottlenecks. As a result of the delay<sup>21</sup> and packet dropping, the entire communication channel becomes unreliable and inefficient. There are several methods for determining whether the information received is correct or incorrect. The first is a hardware-based technique, and the second is a software-based technique. The hardware-based method necessitates advanced computing machines that must be kept in a specialized environment with sophisticated network availability. The software-based design is not very reliable; however, it can identify redundant information generated in the cluster's closely related sensor nodes.

## 1.4 | Motivation

The motivation behind the development of opportunistic energy-efficient clustering algorithms lies in addressing the energy consumption challenges in WSN-IoT network communications and other resource-constrained networks. WSNs consist of a large number of tiny, battery-powered sensors that collect and transmit data from the environment. Prolonging the network's operational lifetime while maintaining its functionality is a critical concern due to the limited energy resources of these sensors. Opportunistic energy-efficient clustering algorithms aim to optimize energy consumption by leveraging the inherent characteristics of the network, communication patterns, and node heterogeneity.

The following section depicts the general structure of the study method: Section 2 discusses the various peer-competing existing protocol and their limitations. Section 3 presents the proposed method and is tested with different scenarios. The result analysis of the proposed and existing protocols is presented in Section 4. Finally, Section 5 concludes the paper and highlights the future scope.

## 2 | LITERATURE REVIEW

This section contains a review of the literature on available routing techniques as presented by various researchers. Because the WSN-IoT applications paves the way for the ultra-modern sophisticated world, the analysis is focused on recent routing technologies for IoT-WSN. So, in this section, an analysis was performed, and the benefits and drawbacks of existing versus proposed schemes were compared. The classification of this literature study is explained in the following subsection.

### 2.1 | Energy-efficient routing in WSN

Chithaluru et al<sup>22</sup> developed an energy-saving routing protocol to extend the lifetime of WSNs, focusing on one dimensional (1-D) queue networks. The resolution on multi-path will be taken by opportunistic routing based on the interval between the node and the BS, sensor node variations, and the average remaining energy. Adaptive ranking-based energy-efficient opportunistic routing (AREOR) claims to keep sensor nodes with low remaining energy and low routing costs. Because AREOR can provide significant improvements, primarily on the 1-D platform, an efficient routing algorithm that can work in a dynamic environment is critical. The proposed protocol can adapt to the changing environment by using a well-organized routing scheme to reduce the effect of the bottleneck, thereby extending the network lifetime.

Awan et al<sup>23</sup> developed a novel cluster-based protocol for cluster centralization in order to provide an optimized WSN. This algorithm is intended for use in a real-time environment. This protocol is the harmony search algorithm (HSA), which is based on harmonies and optimization using a music-based method. In terms of unbalanced overhead scenarios, the HSA by harmony memory considering rate (HMCR) could not outperform. HMCR could try to find a relevant node closer to the CH, but without bottleneck consideration, this HSA is meaningless. In gateway selection, the proposed scheme provides a suitable approach for determining the most preferable path. The clusters will also be controlled using the optimized Steiner tree-based strategy.

Xu et al<sup>24</sup> proposed an energy-stable routing mechanism for efficient data transmission and reception. This protocol is known as a forward aware factor energy balanced routing mechanism (FAF-EBRM). This technique chose the next possible sensor node based on specific parameters such as forward energy density and connection cost sensitivity. In addition, the author has presented a local topology re-configuration mechanism that improves WSN clustering. The performance of this FAF-EBRM is compared to other routing strategies such as LEACH and the energy efficient unequal clustering (EEUC) protocols. FAF-EBRM uses an energy-aware scheme to identify the path based on node weight, and it ignores bottleneck delays. The proposed scheme has the advantage of efficiently identifying the congested route and reconfiguring the broken path. As a result, the likelihood of end-to-end delay will be reduced.

Shahraki et al<sup>25</sup> developed the energy balanced routing protocol (EBRP), a routing protocol for energy balancing. The primary goal of this protocol is to protect low residual energy sensor nodes by forcing data packets on the road to the BS node via the route containing high residual energy sensor nodes. This routing protocol may cause looping problems; to avoid this, enhanced design mechanisms are used. The inability of this protocol to handle end-to-end communication and packet transmission has been overcome by the proposed approach.

Alharbi et al<sup>26</sup> proposed a technique to improve network throughput. This protocol is known as energy efficient opportunistic routing (EEOR). This protocol locates sensor nodes near the BS, and they simply monitor channel communication to begin forwarding packets at any time. This technique sorted the sensor nodes in the forwarding list from lowest to highest priority, and packets received from higher priority nodes are dropped by lower priority nodes. However, in EEOR, the process of creating a priority-based table is complex and time-consuming. Furthermore, EEOR does not address the ambiguous bottleneck issue. The proposed technique efficiently identifies nodes using the optimized Steiner tree algorithm and the minimum spanning tree (MST) technique. Bottlenecks in WSNs can be significantly reduced by the proposed protocol.

Aftab et al<sup>27</sup> presented a novel time resolution tuner (TTR) for time resolution switching. The purpose of this modern tuner technique is to reduce the energy consumption of the micro controller unit in sensor nodes. Time-based packet scheduling in relation to ideal and active sensing conditions may reduce node energy utilization. However, it is critical to consider route traffic and congestion. The energy exploitation problem was effectively addressed by an efficient route selection based on a minimum spanning tree and optimized gateway allocation. The main disadvantage of this TTR is that it only considers time as a factor in energy conservation. The proposed scheme is capable of self-reconfiguring a network in the event of a bottleneck, which is required in energy-aware routing.

Ren et al<sup>28</sup> presented a system-level power conservation scheme for wireless body area networks (WBANs) and determined the optimal distance for sensors to transmit and receive data. The transmission distance and the available energy in the circuit are used to calculate this threshold distance. By analyzing the *dth* threshold, this technique only balances the energy between the circuitry and the data transmission. This scheme's lack of route awareness reduces reliability. To save energy, the proposed protocol addresses this issue through efficient route identification and gateway selection.

Zhu et al<sup>29</sup> introduced a dual-metric K-means (DK-means) algorithm for minimizing correlated data in spatially distributed indoor sensor networks. This DK-means algorithm reduced the redundant information by designating nodes within the cluster to act as representatives and transmit the observed data, thereby saving energy. Although this DK-means technique avoided redundant transmission, it was unable to restructure a broken network in the event of congestion. The proposed method provides a dynamically adoptable self-configurable network and checks for bandwidth to identify the optimal gateway.

Chithaluru et al<sup>30</sup> developed an energy-efficient routing scheme for mobile WSNs. According to the authors, the proposed protocol can dynamically reconfigure in response to sensor node mobility. Furthermore, this protocol selects only sensor nodes to forward packets to the BS, all of which are capable of balancing the energy-saving distribution for the WSN. Because the system did not identify the gateway's bandwidth, a bottleneck occurs if the required bandwidth is greater than the existing one. This technique contains no statements for dealing with the damaged/broken route. By using the MST algorithm, the proposed algorithm can avoid bottlenecks and find the best route.

## 2.2 | Routing schemes for the IoT

Naveen et al<sup>31</sup> presented an efficient mobile gateway key for IoT applications. This study focuses on implementing this protocol in mobile health devices such as patient monitoring systems. The user's or the patient's data is collected independently and forwarded to the intelligent personal assistant (IPA) or the medical center for caretaking. This AMBRO

mobile app is set up to monitor the patient via a personalized mobile gateway. However, for communication, this application used a standard WSN routing protocol. As a result, it encountered common network issues. Through optimal path identification and bandwidth-aware gateway selection, the proposed protocol can improve system efficiency.

Hamidouche et al<sup>32</sup> investigated compressive sensing (CS) perceptions in the IoT environment. Initially, the authors use a compressed sampling technique to analyze the CS theory with a low processing cost. In addition, a novel CS-based approach to calculating, propagating, and accumulating information in the fusion center is proposed. An improved cluster-sparse reconstruction protocol is proposed to reduce energy consumption and improve data reconstruction. So, compression was performed at each access point, and reconstruction was performed based on data dependencies in the fusion cluster. Despite the fact that the node performs the compression technique for accurate information reconstruction, an optimal path and path reconfiguration protocol is required to avoid network issues. Based on MST, the proposed protocol provides an optimal solution for determining the gateway and the least weighted route.

Chithaluru et al<sup>33</sup> proposed an energy-efficient trust derivation protocol to improve WSN security while reducing system overheads. Because WSN performance is sensitive to network overhead, this protocol employs a game theoretic approach to reduce network overhead. The disadvantage of this technique is that it is highly iterative and time-consuming because it requests trust scores from neighbors. The transmission history of a node is used to calculate trust values. The energy expended in these iterative processes should roughly equal the energy saved. Meanwhile, the proposed scheme consumes less time and has less computational complexity because it employs the MST algorithm, which is significantly more efficient than the previously mentioned trust-based approach.

Dev et al<sup>34</sup> illustrated a modern wireless sensor network with energy harvesting (EHWSN) technique; This protocol allowed the network's sensor nodes to mine energy from the surrounding resources. The MAC layer handles all optimizations performed throughout this research. In the event of a broken link, this technique is unable to provide reliable communication. There was also no way to avoid the bottleneck path. The communication links in the proposed technique are formed using the Steiner edge detection scheme. Additionally, the optimized Dijkstra's shortest path algorithm (ODSP) determines the best gateway based on bandwidth. As a result, the proposed technique provides reliable communication in the event of broken links.

Raslan et al<sup>35</sup> developed a new redundancy-based weight election protocol (R-WEP) to allow communication among internet-enabled devices. The case-monitoring weighted sensor networks are used to increase network lifetime based on the weight election protocol; an efficient redundancy mechanism is used in the sensor nodes for re-configuration. This technique chooses the CH and next hop node based on residual energy; the nodes that are exempted as a result of this technique are known as redundant nodes. Because WSNs are intended to be deployed in dynamic environments, the system must be self-configured. The proposed technique clusters the nodes as trees, and the MST algorithm determines the best path based on edge weight. Furthermore, by reducing the number of retransmissions, the network's energy is conserved.

Hafeez et al<sup>36</sup> proposed a standard design concept for the long-term implementation of wireless sensor monitoring systems. This paper takes into account the fundamental requirements of wireless network parameters such as faster deployment, quality of service (QoS), less maintenance, and low cost for all design components. This study only looks at the implementation of a real-time WSN application using a standard field link scheme. The main disadvantage of this concept is that the nodes are not clustered; additionally, the system is not designed to choose an alternate path if the route is damaged. This is overcome by the proposed approach, which generates the route using an edge-weighted tree based on the MST algorithm. In addition, the link is validated for transmission after the successful identification of the optimal gateway.

### 2.3 | Routing for WSN with IoT and gateway

Hussain et al<sup>37</sup> presented a ZigBee and general packet radio service (GPRS)-based IoT gateway system tailored to the needs of telecom operators and various IoT application environments. It implements a model gateway link with the application server, as well as the implementation issues associated with it. This concept makes use of a protocol conversion mechanism in which the gateway repackages the data in accordance with the protocol of the sender's WSN standard. However, this implementation does not specify the system's fault tolerance, which is required to obtain a reliable link for the gateway. The proposed technique defines a proper algorithm for both route selection and gateway identification, and the link should be validated and determined based on the bandwidth.

Campbell et al<sup>38</sup> proposed a rule-based gateway to connect the various IoT protocols and to provide a solution for integration with horizontal IoT services. In IoT application scenarios, this protocol causes contextual fragmentation. In addition, a generic IoT protocol is being developed to document the root cause of the IoT fragmentation problem. This proposed protocol enables the standard message queuing telemetry transport (MQTT) scheme to combine QoS-based parameters and other traffic-aware protocols. Despite the fact that this MQTT-based queuing service provides a solution to some integration issues in horizontal IoT devices. The self-configuration scheme was not used in the gateway. Furthermore, the gateway identification mechanism is inefficient for establishing a dependable link.

## 2.4 | Review on traffic attentive routing protocols

Mijuskovic et al<sup>39</sup> propose a scheduler algorithm for multi-path propagation that uses the multi-path transmission control protocol (MPTCP), which reduces network energy consumption. They analyze multiple application schedules based on the previous history of communications and various radio model energy interfaces using offline Markov decision processing. Even though this technique attempts to reduce network energy consumption, one of the major drawbacks is the lack of a hotspot detection protocol. As a result, the proposed protocol provides traffic-aware routing as well as hotspot identification to stabilize network efficiency.

Yuan et al<sup>40</sup> developed an energy-aware multipath TCP (eMPTCP) design model for mobile devices by analyzing MPTCP energy models in various wireless fidelity (Wi-Fi) and mobile cellular interfaces. This technique identifies the active region in which the MPTCP consumes significantly less energy than the baseline standard protocols. The proposed scheme, which uses self-geographic data to identify the hotspot, provides a more accurate location estimation of energy waste than the eMPTCP model. Furthermore, to handle cluster communication, the proposed technique employs a novel CH selection scheme.

Lenka et al<sup>41</sup> proposed an advanced MPTCP algorithm for improving energy conservation and network performance in mobile WSNs. This protocol considers two different mobile applications: constant interval real-time, constant data size, and continuous data transfer applications. Even though data is transferred in a dynamic path on a regular basis, multi-path propagation is impractical in this approach. Furthermore, because this scheme lacks a traffic detection mechanism, the likelihood of a transmission link disconnection is high. The proposed scheme provides a traffic-aware hot-spot identification protocol that reconfigures the path and reduces the occurrence of bottlenecks using a location-based algorithm.

Hu et al<sup>42</sup> introduced a new multi-path-network utility maximization (MP-NUM) protocol. This protocol is suitable for both multi-hop and single-hop users. A general solution for maximizing multi-path network utility is provided. This study introduces mReno, a novel transmission control protocol (TCP) multipath technique. The MP-NUM protocol determines the routing path based on random time slots, and it has a high delay tolerance. Because of the excessive delay in the retransmission, the sender waits for acknowledgement in this protocol, and the redundant data received at the receiver is dropped. The proposed scheme can optimize enormous energy and delay. Because the proposed scheme relies on real-time network information rather than a random variable, the network's accuracy and reliability would be improved.

Yang et al<sup>43</sup> proposed a typical binary structure with cross-correlation assets for use in a wireless channel. This article also depicts a novel back-off protocol. By removing the need for channel observation, this methodology can reduce the need for energy consumption during data reception. To improve throughput, an additional collision avoidance scheme based on prediction is provided. To avoid the bottleneck, this method uses a probabilistic approach to determine the carrier sensing. It forecasts network state based on node configuration and previous transmission. Despite the fact that it is a deterministic polynomial approach, the significance of energy conservation is insufficient. The proposed method groups nodes based on spanning parameters such as aggregate distance, marginal distance, and BS distances. As a result, it saves more energy. The use of self-physical data improves the accuracy of hot-spot detection, which effectively reduces collisions.

Feng et al<sup>44</sup> proposed a bottleneck prediction scheme in WSN and explained how this could be avoided in the communication channel further. A suitability cost metric function is calculated based on the number of sensor nodes available in the significant routing path. This cost metric function is used to make an estimated prediction of congestion or bottleneck in the WSN. This protocol estimates the cost for all WSN edges and determines the lowest cost to choose the routing path. It does not, however, identify the bottleneck's channel status. A bottleneck can even degrade the efficiency

of a cost-conscious network. As a result, a traffic-aware collision detection scheme for WSN is required, and the proposed routing scheme is derived. It lowers the likelihood of bottlenecks and increases network reliability.

Liu et al.<sup>45</sup> presented a lightweight buffer management scheme for WSN bottleneck prevention. This method can prevent data overflow in the buffer of the middle sensor node. The possibility of congestion occurrence is avoided by optimizing data transmission rates with automatic sensor re-configuration. This protocol reduces the occurrence of congestion by reducing the size of the middleware buffer's stack queue. In addition, if the delay exceeds the threshold, this protocol determines an alternate path. However, an efficient network requires a high throughput rate with low end-to-end latency, which this protocol does not provide. The proposed protocol implements a traffic attention scheme based on hotspot detection. P-TAVR eliminates the need to reduce buffer size or packet transfer rate because it regulates data flow through three phases of control and achieves high throughput with minimal end-to-end delay.

## 2.5 | Review on cluster-based routing

Simiscuka et al.<sup>46</sup> proposed a multicast delivery approach to elucidate surplus energy consumption issues in networks by taking into account node sensing capability. They focused on directional reception antennas to build a multi-hop hierarchy using a two-step reconstruction routing pattern. In this study, two algorithms were used to determine the multicast tree and the power vector of minimum transmission. One of the most significant drawbacks of this technique is that it does not use a scheduling policy to regulate data packets over the channel.

Abd-Elmagid et al.<sup>47</sup> proposed a three-pronged approach to power devastation computing. Energy flow analysis (EFA), input-output analysis (IOA), and ecological network analysis (ENA) were the approaches used. Their relative importance and decision-making capability were investigated in the context of central power adoption. This protocol is intended to investigate and examine public power and economic investigations, with the energy flow reasoning managing the scrutiny of initial and final power usage. This study takes into account the energy flow in both inbound and outbound urban power devastation scenarios. ENA was used to discover the driving economic sectors of energy consumption, as well as the control/dependence relationships between sectors. To consume energy, the proposed tree-based clustering technique and typical scheduling algorithm take into account the important energy-related parameters.

According to Zou et al.,<sup>48</sup> the cutting edge in the field of WSN is defined by low energy consumption. This investigation discusses the advancements based on the WSN lineaments. In addition, network topologies, power sources, and management level summaries are examined. This investigation and examination are limited to the hardware devices and antennas that use WSN energy in precision agriculture. Because the routing protocol is one of the most important factors influencing a network's energy consumption, it is necessary to consider the network's routing schemes and scheduling policies. In a typical WSN, the proposed scheme derives a trust-based scheduling approach as well as an efficient clustering scheme, and energy consumption has been analyzed in relation to various parameters.

Yang et al.<sup>49</sup> proposed a modern data acquisition scheme that is reliable and an energy-efficient routing (RMER) that addresses the prerequisites of stability and power efficiency. Fewer sensing nodes are preferred in the coverage region, while a large number of sensor nodes are chosen in and around the non-coverage region, reducing sensor node power consumption and maximizing network lifetime. The disadvantage of this RMER technique is that it favors sensor nodes located outside of the coverage region. As a result, the source/sender uses a lot of energy to send data over a specific node, which reduces network lifetime. The proposed scheduling policy and optimal clustering scheme take energy-related parameters into account. As a result, it consumes less energy and increases network lifetime. Table 1 tabulated the performance evaluation of proposed and peer-existing methods.

**TABLE 1** Performance evaluation of proposed and peer-existing methods.

Method	ARE	Throughput	PDR	End-to-end delay
Awan et al. <sup>23</sup>	Moderately less	Less	Good	Moderate
Zhu et al. <sup>29</sup>	Moderately less	High	Good	Good
Chithaluru et al. <sup>33</sup>	Less	High	Good	Good
Feng et al. <sup>44</sup>	Less	High	Moderate	Good
Proposed	High	High	Very good	Very good

### 3 | PROPOSED METHODOLOGY

Congestion control in WSNs is a significant challenge and critical issue that can lead to inherent resource constraints, disrupted communication, and network instability. Congestion has a significant impact on QoS, PDR, end-to-end delay, and energy consumption; therefore, it must be managed. In WSNs, parameters such as collision, buffer overflow, channel constraints, and transmission rate cause congestion. WSN congestion can be alleviated in two ways: by reducing data traffic or by increasing network resources. When there is congestion, packets are not properly received between the intermediate nodes, making appropriate routing to the BS node impossible. The proposed method, which employs the optimized Steiner minimum tree-based routing, can be implemented in different configurations. In addition, all of these configurations necessitate optimal interconnections for a predefined set of objects and an objective function. The Steiner tree problem in graphs is a well-known variant that often gets confused with the Steiner tree problem. The Steiner tree problem in graphs necessitates the construction of a minimum-weight tree that contains all of the terminals for a given undirected graph with non-negative edge weights and a subset of vertices known as terminals. Other well-known variants include the optimized Euclidean Steiner tree problem and the rectilinear minimum Steiner tree problem.

When there are multiple paths to take from a source node to a destination node, an optimized Steiner tree reduces the number of forwarders and builds multiple trees in parallel with fewer common nodes; that is, multiple trees are built from the source node via different neighboring nodes, with fewer common nodes among the trees built from these neighboring nodes. According to its definition, it reduces the number of nodes and links used to construct a delivery tree. As a result, it is extremely useful for representing multicast routing solutions.

#### 3.1 | Network model

The network model used is IEEE 802.11ah, which is a subset of the IEEE 802.11 standard with some physical and MAC layer revisions. IEEE 802.11ah allows networks to operate below 1 GHz, with a minimum data rate of 100 Kbps and a single-hop transmission range of up to 500 m not including white space bands. IEEE 802.15.4 can handle data rates up to 500 Kbps in an unlicensed 4.8 GHz band, but 802.11ah can handle data rates ranging from 710 Kbps to 6.1 Mbps in a sub-2 GHz license-exempt band with a larger coverage area. By providing a hierarchical network organization that improves simplicity and stability, the 802.11ah can associate with a greater number of stations. Furthermore, IEEE 802.11ah uses a fast association technique to avoid collisions. However, 802.15.4 can collaborate with 48,000 devices, putting an undue strain on the BS.

#### 3.2 | ODSP algorithm

As a new shortest-path algorithm, the ODSP algorithm is proposed. Instead of a single parameter, this algorithm used multiple parameters to find the shortest valid path. The efficiency of the ODSP algorithm was assessed in terms of the shortest path by measuring its nodes and time complexity. Individual networks are extremely vulnerable to node failures due to the gateway disconnection issue. However, a hybrid network with multiple gateways reduces the issue of node failure. One of the dense IEEE 802.11ah protocols is used here to achieve energy-efficient communication. Figure 3 represents the classical design of a dense IEEE 802.11ah access point. The proposed architecture supports both wired and wireless structures. The WSN-IoT middleware and the gateway are linked by a wired connection. The gateway's connection to the WSN-IoT middleware is wired, but the sensor nodes and the gateway communicate via a wireless link.

This section describes the working principles, system architecture, and critical steps involved in the proposed technique. Figure 4 depicts the proposed scheme's block diagram. The first step is to create the IoT architecture. The nodes in the network are then clustered using the Steiner minimum tree. The gateway is then adopted between the controller and the radio network in the following step. The gateway receives and saves one copy of the information from the CH before transmitting it to the Internet for proper processing, depending on the application's requirements. The final step is link validation, which requests available bandwidth to prevent packet drops.

As shown below, there are three IoT architecture tiers.

- **Client tier:** The client-level structure acts as a bridge between the IoT device and the network. It makes services more accessible to the end user. The services may differ depending on the data collected by the IoT sensor. At the

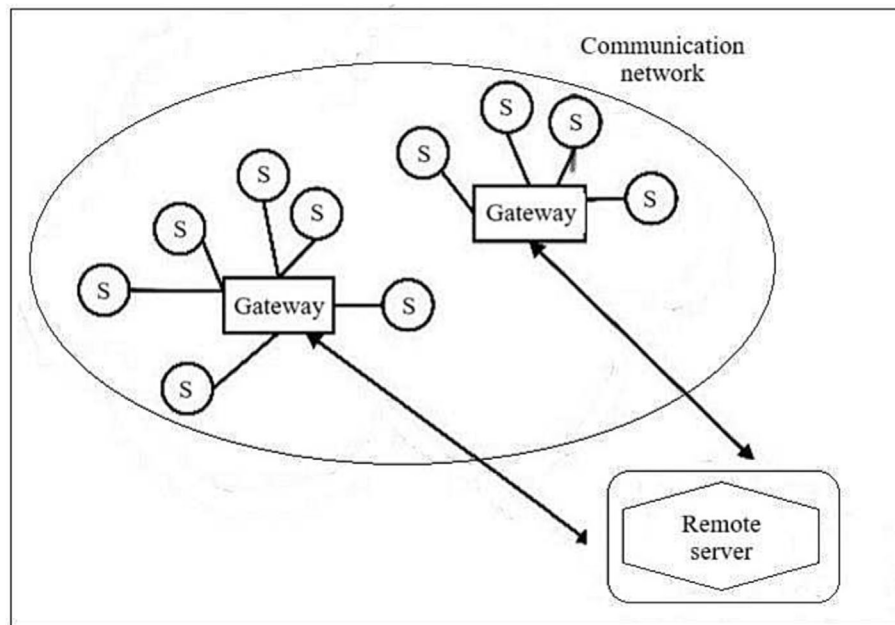


FIGURE 3 Structure of dense network IEEE 802.11ah.

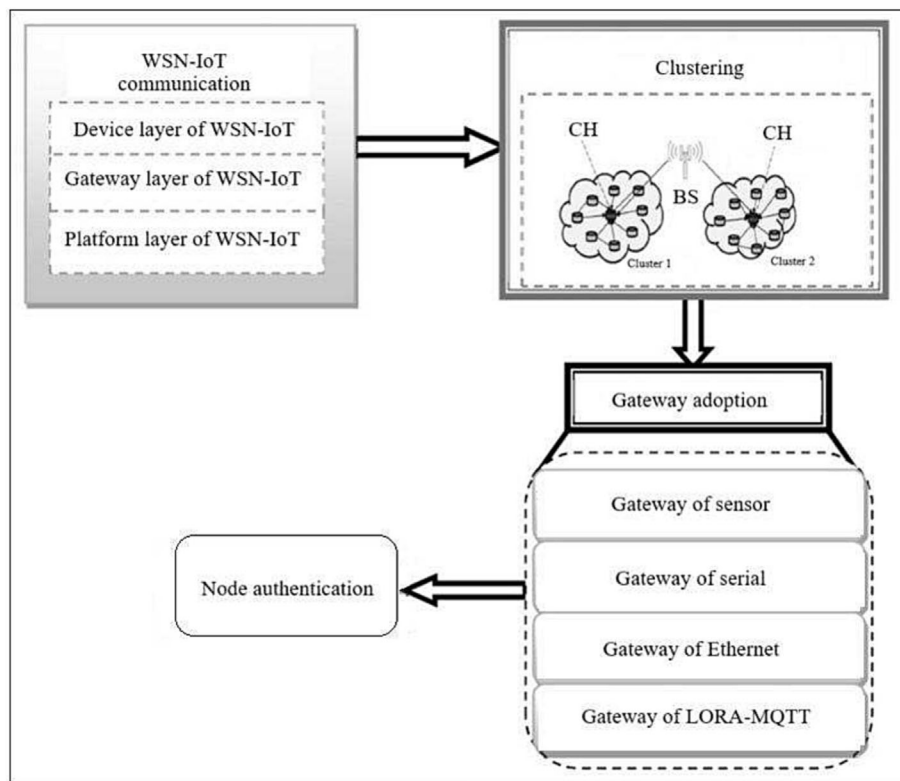


FIGURE 4 Functional diagram of proposed scheme.

client level, an application tier is created that defines each IoT sensor. A computer or mobile device browser, for example, can access the application tier by executing protocols such as HTTP, FTP, and SMTP.

- **Server tier:** At this level, data processing and networking are managed. Pre-shared keys and passwords are used to authenticate data received from clients. The validated data is then transmitted to the network layer via a wired or wireless medium. The network layer is responsible for transporting sensed data between IoT Sensors and the network.
- **Operator connection:** The operators, who are typically service providers, are in charge of a client's validation and authentication. Based on the client's request, the operator link redefines the service. It has the ability to add or



remove clients/devices from the network. It also releases patches for the application tier in order to upgrade or correct any errors.

Because this is the basis of the feasibility criterion, consistency results in superior design performance. Figure 5 depicts the flow chart of the proposed routing scheme architecture.

### 3.3 | Clustering

The BS always generates an aggregated value for the end-users, and aggregating the data to be forwarded can also help to reduce transmission overhead and energy consumption. Nodes in the network can be accommodated in small groups known as clusters to support data aggregation. Clustering is the process of assembling nodes into groups based on a

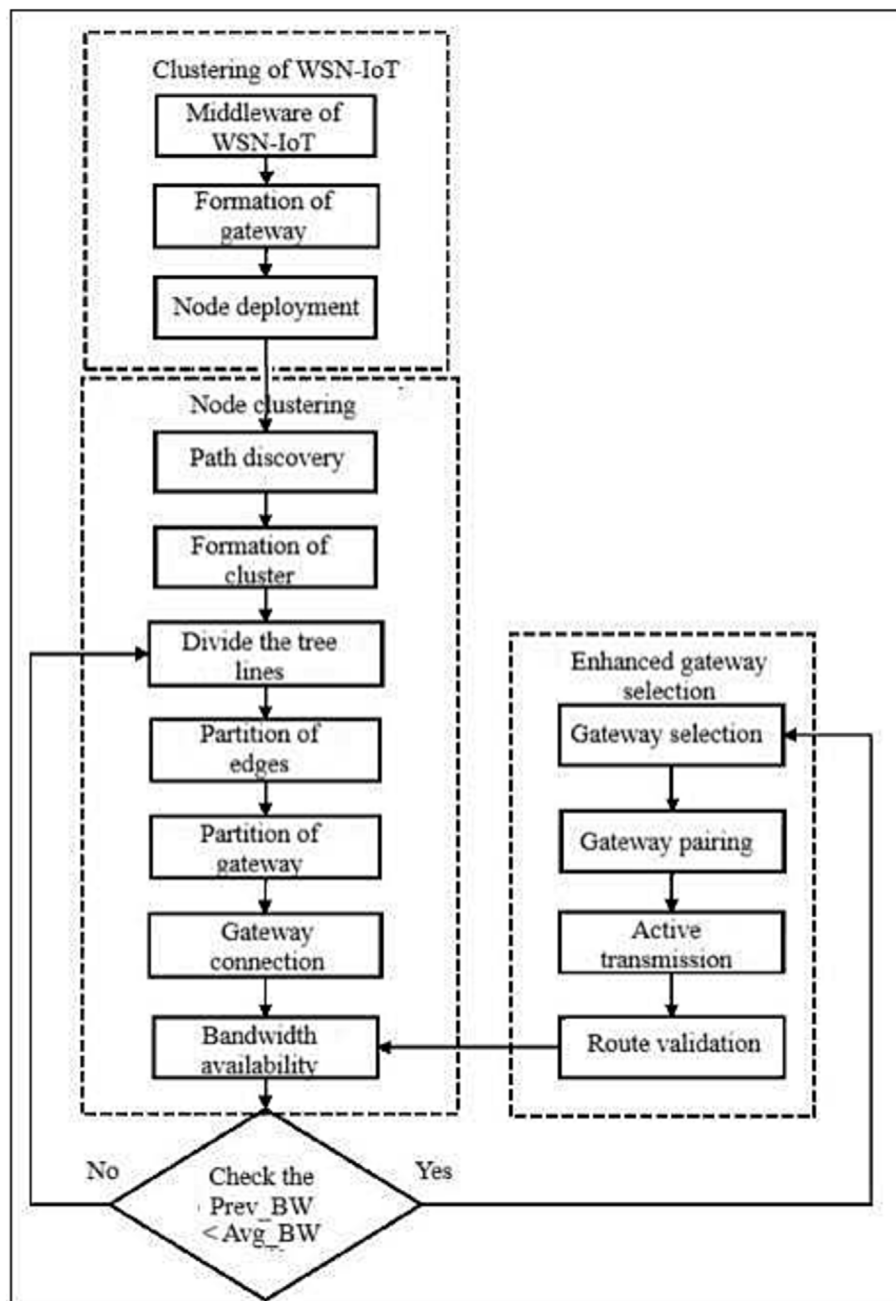


FIGURE 5 Architecture of proposed routing scheme.

mechanism. Clustering has been shown to improve network lifetime, a key metric for assessing a WSN's performance. Clustering is used to enhance energy efficiency. Figure 6 represents the clustering of nodes.

The information obtained from neighbor nodes is used to construct a graph. The graph has vertices named  $v_1, v_2, \dots, v_n$  and an input root named  $\rho$ . Then, another graph is created with the vertex, edge, and edge interval being  $v_\rho$ ,  $\epsilon_\rho$ , and  $\kappa_\rho$ , respectively. The apex, boundary, and station sets for a weighted subgraph are created from the previous graph. Based on the cost of their boundaries, these cluster participants are combined into a cluster. The stages involved in the formation of cluster members are depicted in Algorithm 1,

---

### Algorithm 1 Pseudo code of clustering

---

- 1: **Input:** Generation of minimum weighted graph  $\xi_\rho$ , Where  $\xi$  represents initial Graph, Vertex  $v_\rho$ , edge  $\epsilon_\rho$ , cost  $\kappa_\rho$ .
  - 2: **Output:** Best route  $\tau$ .
  - 3: **Procedure** Clustering
  - 4: **begin:**
  - 5:  $\eta = 0$  ▷ Initially no members
  - 6: Create a subgraph  $\eta = \eta + 1$  ▷ Increase the cluster members
  - 7: Set maximum number of Cluster  $\mu_k$  is a subset of the root node  $\{\rho\}$ .
  - 8: Highest cluster module value as the subset  $\sigma$ .
  - 9: Minimum congestion path  $v_0, v_1, \dots, v_e: v_e$  and  $\mu_k$ , where  $v_e$  as cluster  $\mu_k$ ,  $f(v_0)$  as element of  $f(\mu_k)$ ,  $v_i$  as vector set  $v_\rho$  and  $f(v_j)$  as subset of  $f(\mu_k)$ .
  - 10:  $\eta = \eta + 1$  Clustering of  $\{u, v\}$ . ▷ Increases the members count
  - 11: Set  $P_i$  as  $\sigma\sigma$  represents the ratio of maximum subset clustering value of  $(v_i, v_j)$  to the actual value of  $(v_i, v_j)$ . where,  $(v_i, v_j)$  lies within the boundary of cluster  $\chi$ .
  - 12: Select path  $P$  from  $\sigma$ , such that the members in the cluster  $\mu_k$  is the combination of  $\mu_k$ .
  - 13: Construct graph  $\xi\{v, e, c'\}$  as apex, boundary and terminal.
  - 14: Set average of  $\chi(u)$  and  $\chi(v)$  as cost ▷ for all the boundaries in the graph  $\xi'$
  - 15: Select optimal path  $\tau$ . ▷ the known approximation technique  $\xi'$  and  $U$
  - 16: Remove the unwanted nodes. ▷ from the cluster members  $\xi_k$
  - 17: **end Procedure**
- 

The information is properly routed through an optimal gateway after the cluster is formed with CHs and cluster members.

### 3.4 | Optimal gateway adoption

The gateway is primarily made up of manually entered router internet protocol (IP) addresses in the host. This is the most commonly used method because it is simple. However, if two or more routers are connected to the same subnet,

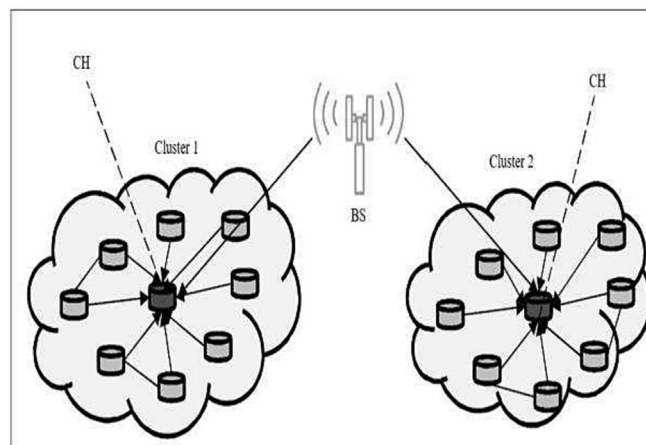


FIGURE 6 Clustering of nodes.

the network administrator must decide which of these routers should receive the message. To resolve this issue, a post office protocol (PoP) redirect message is available. The default gateway examines the destination host address of the received message. If it determines that the message should be routed via another router, it will send a redirect message to the source host with the IP address of the selected router. When the source host receives this message, it saves both the destination host address and the second router address in order to send messages to this host via the second router.

- **Sensor gateway**—It is an all-in-one board that, depending on the software configuration, can function as a serial gateway or an ethernet gateway. With a plug-and-play option, this is simple to use.
- **Serial gateway**—This gateway communicates with the controller directly via an available universal serial bus (USB) port. Because the serial gateway relies on available USB ports, it should be located near the controller.
- **Gateway of Ethernet**—This gateway connects to the Ethernet, and the controller allows for greater placement flexibility than the serial gateway. Furthermore, the gateway can be placed in a central location on the radio network.
- **Long range-message queuing telemetry transport (LoRa-MQTT) gateway**—This gateway also connects to the ethernet network and exposes a LoRa-MQTT broker that can be used with controllers that support MQTT, such as an open home automation bus.

Before constructing a gateway, choose an option supported by the controller. Algorithm 2 determines the best gateway.

---

#### Algorithm 2 Pseudo code of optimal gateway adoption

---

```

1: Procedure Optimal gateway
2: begin:
3: for Every vertex  $v$  belongs to the vertex do
4:   Set  $v$ , bottleneck cost  $\omega_g$ , and the source to the gateway  $\Omega$ .
5: end for
6: for Vertex  $v$  to be  $+\infty$  do
7:   Set lowest bottleneck cost  $\omega_g[s]$ , where  $s$  represents the set of nodes in the lowest bottleneck path from source to
   the gateway  $\Omega$ . ▷ source is assumed to be  $-\infty$ 
8: end for
9: if  $[\alpha] \neq \phi$  then
10:  Set the nodes;
11: else
12:  Mine minimum Value bottleneck value  $\phi$  from  $[\alpha]$ .
13: end if
14: Assign  $[\alpha] \rightarrow u$ .
15:  $\theta$  as lowest destination bottleneck path of  $u$ .
16: while Vertex  $(u, v)$  belongs to the boundary  $\epsilon$  of the cluster  $\chi$  do
17:   if lowest bottleneck weight  $\omega_g[s] >$  maximum value of the set  $\omega_g[u]$  and  $\omega_g[u, v]$  then
18:     Assign maximum value of the set  $\omega_g[u]$  and  $\omega_g[u, v]$  to  $\omega_g[v]$  minimum bottleneck  $u$  previous set  $[v]$ .
19:   end if
20: end while
21: end Procedure

```

---

### 3.5 | Route validation

When the occurrence of a bottleneck in the gateway increases, the likelihood of a packet drop increases. The link validation stage is included to determine the available bandwidth and avoid packet loss. Figure 7 depicts bandwidth-based transmission.

Because the bandwidth required for transmission is less than the available bandwidth of the optimal gateway  $R_2$ , the link between  $R_1$  and BS is validated, and the packets are transmitted. As a result, if the required bandwidth for transmission is less than the bandwidth available in the channel, the gateway link will be authenticated and selected

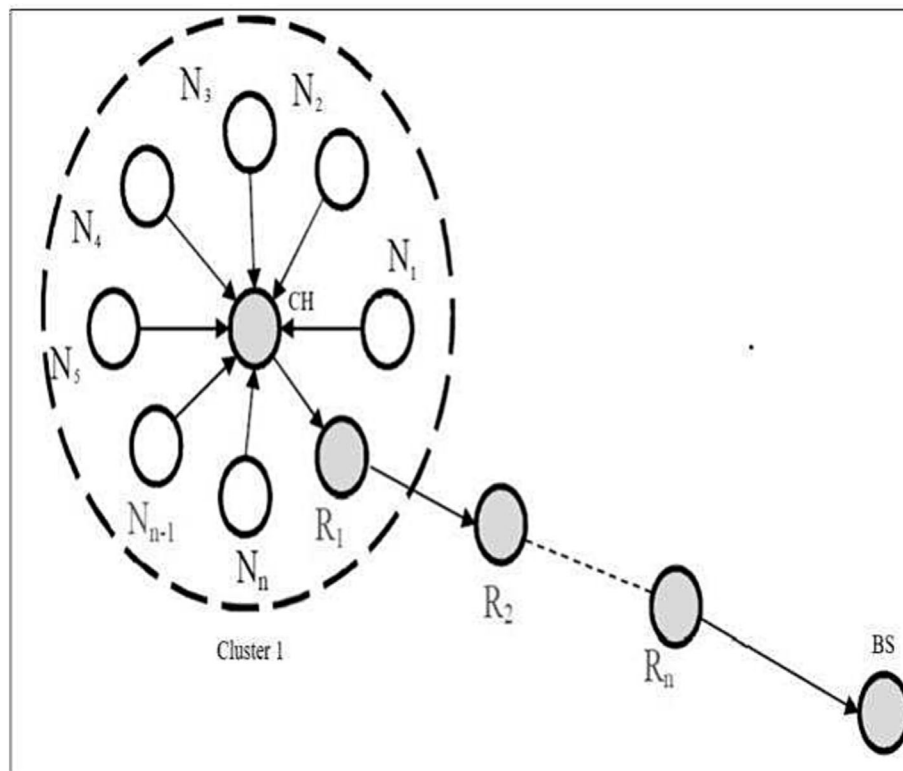


FIGURE 7 Bandwidth-based transmission.

for packet transmission. For higher bandwidth transmissions, the protocol will identify another gateway, and the process will be repeated until an optimal gateway link is authenticated.

## 4 | RESULTS AND DISCUSSION

This section compares the simulation results for the proposed technique to the results of other techniques. The proposed protocol's performance can be measured by comparing it to existing schemes such as AREOR,<sup>22</sup> adaptive ranking-based improved opportunistic routing (ARIOR),<sup>30</sup> and adaptive ranking fuzzy-based energy-efficient opportunistic routing (ARFOR).<sup>33</sup> Network Simulator 2.34 (NS2.34) was the simulation tool used for the performance analysis.

### 4.1 | Network parameters

The network parameters developed in this concept are based on a typical IEEE 802.11ah WPAN. The network's dimensions were  $1000\text{m} \times 1000\text{m}$ . The deployment of sensor nodes was distributed at random. The spanning distance between the nodes ranged from 25 m to 50 m. In this network, 100 sensor nodes with an average energy of 1.5 J each were distributed uniformly. One BS node served this network. These sensor nodes had an observing interval of 10 ms. Each sensor node was capable of transmitting packets of up to 512-bits per second (bps). The available bandwidth for this network model was 40 MHz. The total simulation time was 1500 s. User datagram protocol/constant bit rate was the application type. Table 2 lists the network parameters used in the simulation, which were derived from literary works.

### 4.2 | ARE

The ARE parameter determines which sensor nodes in the routing path have the most remaining energy. Sensor nodes with the highest ARE can participate in communication for a longer period of time. As a result, a system with a high

TABLE 2 Network parameters.

Parameter	Units
Distribution area	1000m × 1000m
Nature of distribution	Uniform
Nodes span	25 to 50 m
Total nodes	100
Source node	1
Initial node energy	1.5 J
BS	1
Maximum coverage	100 m
Bandwidth	40 MHz
Transfer rate	10 packet/s
Packet size	512-bits
Simulation time	1500 s

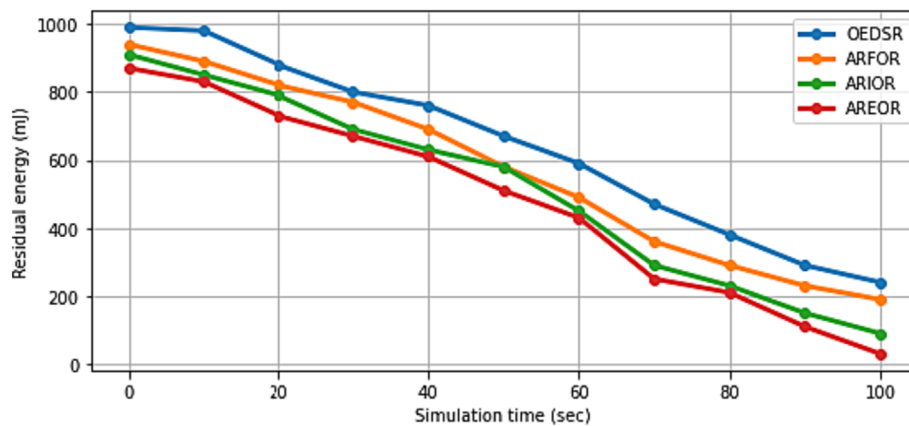


FIGURE 8 Performance analysis of ARE over proposed versus peer routing protocols.

ARE is considered an energy-efficient wireless network. Figure 8 depicts the ARE analysis for the proposed technique and its comparison with the AREOR, ARIOR, and ARFOR protocols. When compared to peer existing routing protocols, the proposed protocol demonstrated the highest ARE.

At simulation time 100 s, the energy level in the existing AREOR technique was 326 mJ. However, the proposed efficient clustering and optimized Steiner tree routing algorithm achieved higher residual energy, resulting in an energy level of 248 mJ at 100 s. The results show that the proposed protocol resulted in a 53.7% decrease in energy consumption.

### 4.3 | PDR

PDR is calculated as the ratio of total packets acquired at the terminal node to total packets transmitted from the source node. As illustrated by the graphical representation in Figure 9, while parallel connections are augmented, the proposed scheme achieves high PDR when compared to peer existing routing protocols.

In the existing protocols, the PDR was high in the AREOR, with PDR values for the least and most parallel links of 87.7% and 51.8%, respectively. Following the implementation of the proposed protocol, the PDR increased to 85.1% and 39.32%, respectively. As a result, the proposed scheme increased PDR values by 2.34% and 11.67% for the least and most parallel links, respectively.

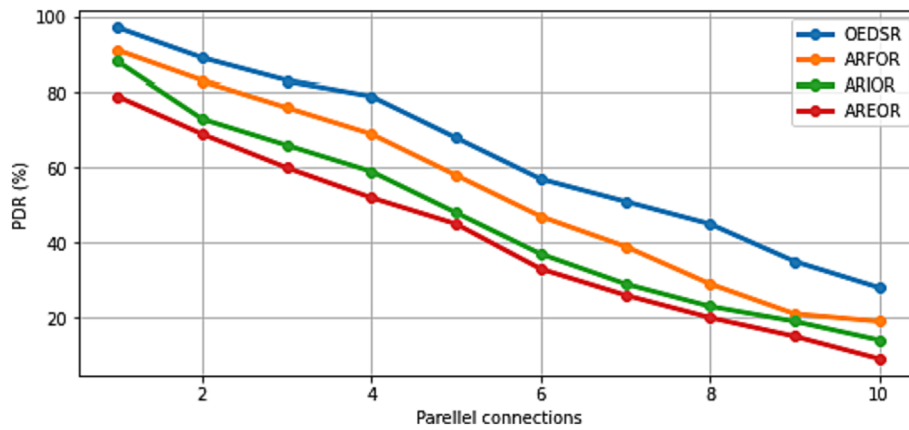


FIGURE 9 Performance analysis of PDR over proposed versus peer routing protocols.

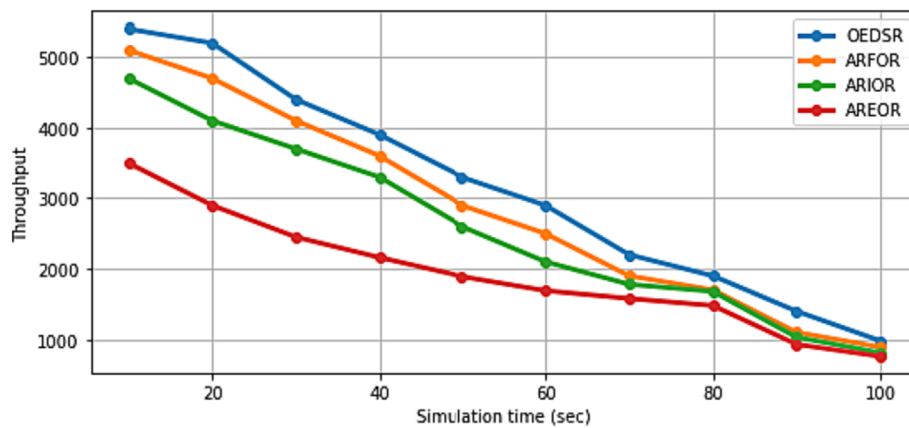


FIGURE 10 Performance analysis of throughput over proposed versus peer routing protocols.

#### 4.4 | Throughput

The ratio of data packets acquired at the terminal node to data packets forwarded by the origin node is known as throughput.

Figure 10 depicts the proposed protocol's throughput analysis and comparison with other available techniques. When considering the throughput rate, it can be seen that the proposed scheme outperforms the other protocols. Among the existing protocols, AREOR has the highest throughput rate between simulation times 0 and 100, achieving an average of 3217 bps. The proposed protocol, on the other hand, achieved an average throughput of 4237 bps. As a result, it is clear that the proposed protocol resulted in a 35.6% increase in average throughput.

#### 4.5 | End-to-end delay

End-to-end delay can be caused by a variety of factors, including optimal path selection, queue length, and communication period. Figure 11 depicts the simulation analysis and comparison of the end-to-end delay based on the number of nodes. When compared to the existing protocols, the proposed protocol has the least amount of delay.

The AREOR experienced significantly less delay in the existing protocols, with the end-to-end delay for the minimum and maximum number of nodes being 42.7 ms and 187.8 ms, respectively. The proposed protocol's implementation of the optimized Steiner tree routing algorithm reduced the end-to-end delay for the minimum and maximum number of nodes to 31.87 ms and 148.9 ms, respectively. According to the results, the proposed technique reduced the end-to-end delay by up to 41.8% and 10.9% for the minimum and maximum number of nodes, respectively.

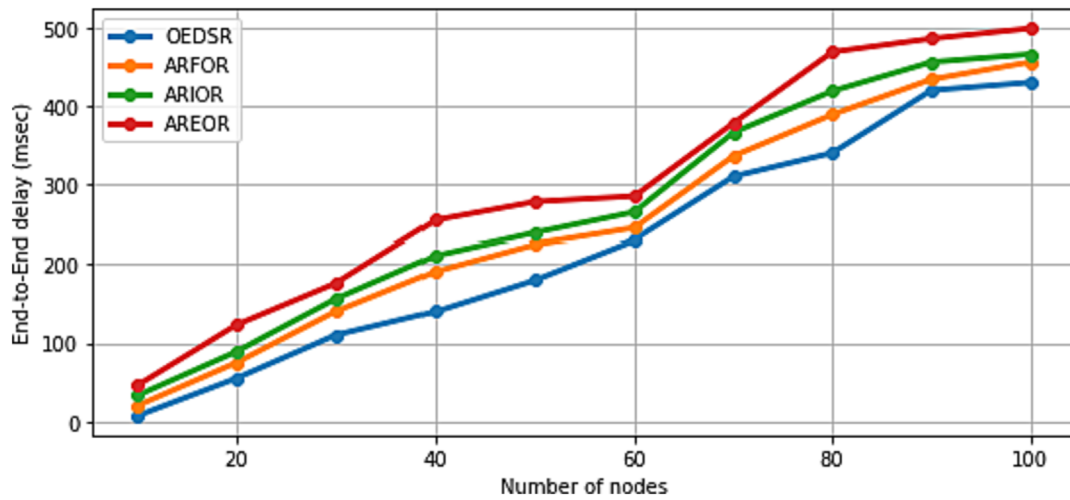


FIGURE 11 Performance analysis of end-to-end delay over proposed versus peer routing protocols.

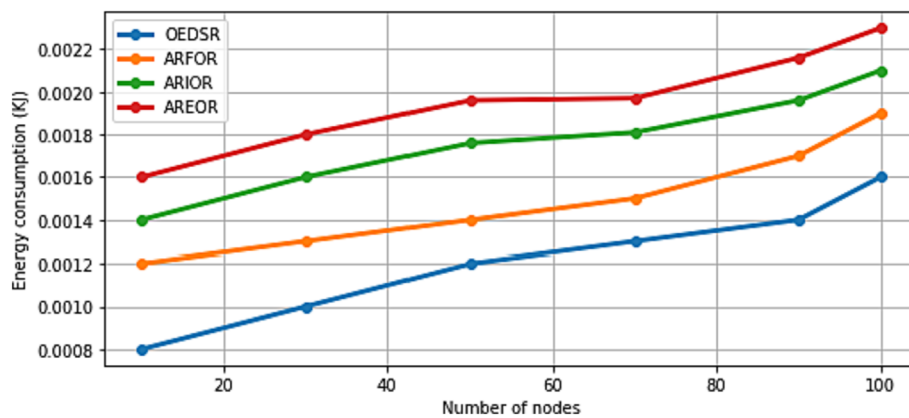


FIGURE 12 Performance analysis of energy consumption over proposed versus peer routing protocols.

#### 4.6 | Energy consumption

Energy consumption includes total energy consumed during data packet transmission and reception, control packet generation, and energy consumed by a WSN node in an idle state. Energy consumption is an important factor because it influences a number of other parameters that are directly related to the performance and operational viability of IoT networks.

Figure 12 shows that OEDSR consumes approximately 47%, 41%, and 38% less energy than the AREOR, ARIOR, and ARFOR protocols, respectively. Power consumption is kept to a minimum even at the maximum density of 100 nodes. When using the fewest possible nodes in a cluster, the proposed method uses approximately 38% less power than its predecessor.

#### 4.7 | Network energy efficiency

Energy efficiency is a factor that defines the average energy consumed per unit bit in a WSN and is expressed in joules/bits. In other words, energy efficiency is a measure of the energy cost per bit of data transmitted in a network. It is another important factor that determines the cost of energy efficiency for a routing protocol. Figure 13 shows that the proposed protocol consumes 51%, 42%, and 35% less energy per bit than the AREOR, ARIOR, and ARFOR protocols, respectively, even at 100 nodes. As previously discussed, the best energy efficiency proposed is due to its low energy consumption and high PDR among its peers.

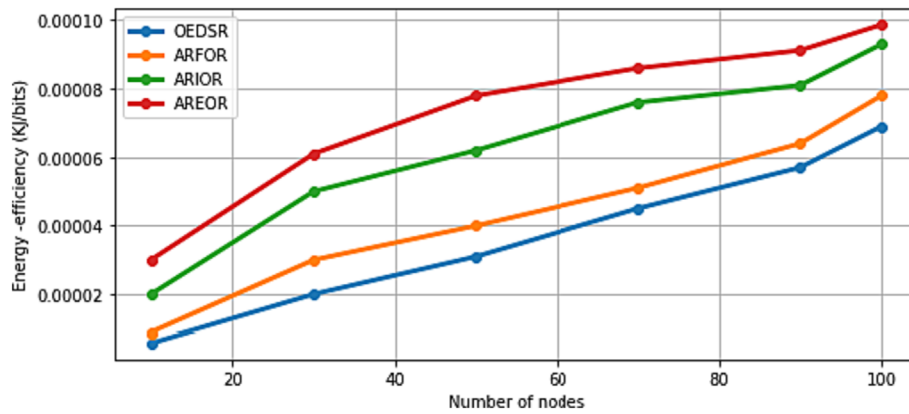


FIGURE 13 Performance analysis of energy efficiency over proposed versus peer routing protocols.

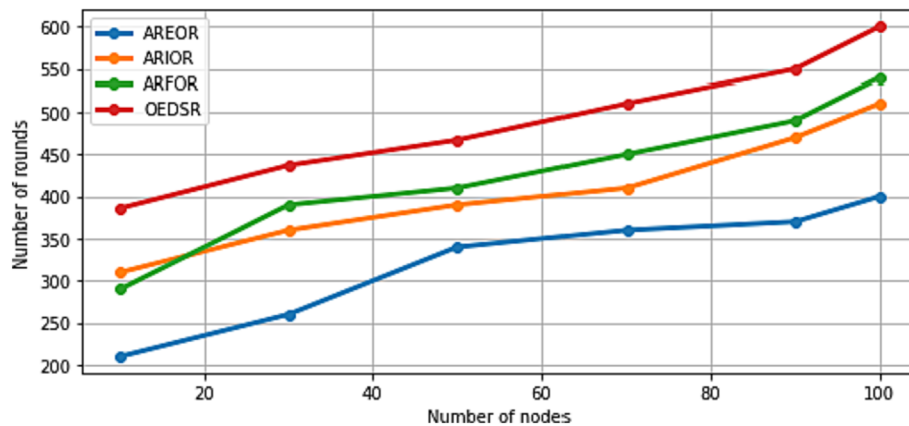


FIGURE 14 Performance analysis of network lifetime over proposed versus peer routing protocols.

## 4.8 | Network lifetime

Network life is defined as the number of rounds that protocol support receives in an existing network before all nodes are exhausted, or as the amount of time that a WSN is fully operational. As a result, it measures the amount of time the network is operational.

Figure 14 shows that the proposed method supports the most rounds. It has roughly 21% more rounds than its nearest competitor, ARFOR. In this scenario, OEDSR gains 81%, 45%, and 21% more rounds than AREOR, ARIOR, and ARFOR, respectively.

## 5 | CONCLUSION AND FUTURE SCOPE

An optimized OEDSR algorithm is proposed in this paper. First, the mobility and energy factors of the sensor nodes are measured, and an optimal route to the BS is determined based on the computation of both values. The number of connections is then reduced by creating a cluster using a hierarchical tree design to reduce power consumption in the WSN. The graph theory approach achieves sensor node clustering via our proposed Steiner tree algorithm; additionally, the central sensor node of a cluster is chosen as the CH. The proposed protocol was compared to the existing protocols, namely AREOR, ARIOR, and ARFOR. PDR, throughput, end-to-end delay, and AER are network parameters that are determined. The measured results show that the proposed protocol outperforms the existing peer-to-peer routing protocols. The reasons for the proposed technique's high performance are discussed further below. The proposed protocol is adaptable to changing environments. As a result, it reduces the impact of bottlenecks and increases network lifetime. It can efficiently identify the congested route and reconfigure the broken path. As a result, the likelihood of end-to-end delay is reduced. The communication links in the proposed technique are formed using the optimized Steiner edge detection scheme. Furthermore, the best gateway for bandwidth is determined. As a result, the proposed technique



provides reliable communication in the event of broken links. Optimized data management should be provided for resource-constrained WSNs in the future.

## ACKNOWLEDGMENTS

Open access publishing facilitated by University of Wollongong, as part of the Wiley - University of Wollongong agreement via the Council of Australian University Librarians.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

## ORCID

Premkumar Chithaluru  <https://orcid.org/0000-0002-1174-1731>

Manoj Kumar  <https://orcid.org/0000-0001-5113-0639>

## REFERENCES

- Chithaluru P, Singh A, Mahmoud MS, et al. An enhanced opportunistic rank-based parent node selection for sustainable & smart IoT networks. *Sustain Energy Technol Assess*. 2023;56:103079.
- Yadav A, Ali Albahar M, Chithaluru P, et al. Hybridizing artificial intelligence algorithms for forecasting of sediment load with Multi-Objective optimization. *Water*. 2023;15(3):522.
- Chithaluru P, Al-Turjman F, Kumar M, Stephan T. Energy-balanced neuro-fuzzy dynamic clustering scheme for green & sustainable IoT based smart cities. *Sustain Cities Soc*. 2023;90:104366.
- Chithaluru P, Fadi AT, Kumar M, Stephan T. Computational intelligence inspired adaptive opportunistic clustering approach for industrial IoT networks. *IEEE Int Things J*. 2023; 2023.
- Pradhan AK, Das K, Mishra D, Chithaluru P. Optimizing CNN–LSTM hybrid classifier using HCA for biomedical image classification. *Expert Syst*. 2023;40(5):e13235.
- Joshi D, Ali Albahar M, Chithaluru P, Singh A, Yadav A, Miro Y. A novel approach to integrating uncertainty into a push re-label network flow algorithm for pit optimization. *Mathematics*. 2022;10(24):4803.
- Chithaluru P, Jena L, Singh D, Ravi Teja KMV. An adaptive fuzzy-based clustering model for healthcare wireless sensor networks. In: *Ambient Intelligence in Health Care: Proceedings of ICAIHC 2022*. Springer Nature Singapore; 2022:1-10.
- Yadav A, Chithaluru P, Singh A, et al. An enhanced feed-forward back propagation Levenberg-Marquardt algorithm for suspended sediment yield modeling. *Water*. 2022;14(22):3714.
- Joshi D, Chithaluru P, Singh A, et al. An optimized open pit mine application for limestone quarry production scheduling to maximize net present value. *Mathematics*. 2022;10(21):4140.
- Chithaluru P, Stephan T, Kumar M, Nayyar A. An enhanced energy-efficient fuzzy-based cognitive radio scheme for IoT. *Neural Comput Appl*. 2022;34(21):19193-19215.
- Jain A, Singh J, Kumar S, Florin-Emilian T, Traian Candin M, Chithaluru P. Improved recurrent neural network schema for validating digital signatures in VANET. *Mathematics*. 2022;10(20):3895.
- Lee W, Jung BC, Lee H. DeCoNet: density clustering-based base station control for energy-efficient cellular IoT networks. *IEEE Access*. 2020;8:120881-120891.
- Shuja J, Humayun MA, Alasmary W, Sinky H, Alanazi E, Khan MK. Resource efficient geo-textual hierarchical clustering framework for social IoT applications. *IEEE Sensors J*. 2021;21(22):25114-25122.
- Zhang Q, Zhu C, Yang LT, Chen Z, Zhao L, Li P. An incremental CFS algorithm for clustering large data in industrial internet of things. *IEEE Trans Industr Inform*. 2017;13(3):1193-1201.
- Hassan AAH, Shah WM, Habeb AHH, Othman MFI, Al-Mhiqani MN. An improved energy-efficient clustering protocol to prolong the lifetime of the WSN-based IoT. *IEEE Access*. 2020;8:200500-200517.
- Yuan B, Lin C, Zhao H, et al. Secure data transportation with software-defined networking and k-n secret sharing for high-confidence IoT services. *IEEE Int Things J*. 2020;7(9):7967-7981.
- Kashef R. Enhancing the role of large-scale recommendation systems in the IoT context. *IEEE Access*. 2020;8:178248-178257.
- Xiong J, Ren J, Chen L, et al. Enhancing privacy and availability for data clustering in intelligent electrical service of IoT. *IEEE Int Things J*. 2018;6(2):1530-1540.
- Sivanathan A, Gharakheili HH, Sivaraman V. Detecting behavioral change of IoT devices using clustering-based network traffic modeling. *IEEE Int Things J*. 2020;7(8):7295-7309.
- Shao X, Yang C, Chen D, Zhao N, Yu FR. Dynamic IoT device clustering and energy management with hybrid NOMA systems. *IEEE Trans Industr Inform*. 2018;14(10):4622-4630.
- Lyu L, Jin J, Rajasegarar S, He X, Palaniswami M. Fog-empowered anomaly detection in IoT using hyperellipsoidal clustering. *IEEE Int Things J*. 2017;4(5):1174-1184.
- Chithaluru P, Tiwari R, Kumar K. AREOR—adaptive ranking based energy efficient opportunistic routing scheme in wireless sensor network. *Comput Netw*. 2019;162:106863.

23. Awan KA, Din IU, Almogren A, Guizani M, Khan S. StabTrust—a stable and centralized trust-based clustering mechanism for IoT enabled vehicular ad-hoc networks. *IEEE Access*. 2020;8:21159-21177.
24. Xu L, Collier R, O'Hare GM. A survey of clustering techniques in WSNs and consideration of the challenges of applying such to 5G IoT scenarios. *IEEE Int Things J*. 2017;4(5):1229-1249.
25. Shahraki A, Taherkordi A, Haugen Ø, Eliassen F. A survey and future directions on clustering: from WSNs to IoT and modern networking paradigms. *IEEE Trans Netw Service Manag*. 2020;18(2):2242-2274.
26. Alharbi MA, Kolberg M. Improved unequal-clustering and routing protocol. *IEEE Sensors J*. 2021;21(20):23711-23721.
27. Aftab F, Khan A, Zhang Z. Hybrid self-organized clustering scheme for drone based cognitive Internet of Things. *IEEE Access*. 2019;7:56217-56227.
28. Ren Z, Mukherjee M, Lloret J, Venu P. Multiple kernel driven clustering with locally consistent and selfish graph in industrial IoT. *IEEE Trans Industr Inform*. 2020;17(4):2956-2963.
29. Zhu B, Bedeer E, Nguyen HH, Barton R, Henry J. Joint cluster head selection and trajectory planning in UAV-aided IoT networks by reinforcement learning with sequential model. *IEEE Int Things J*. 2021;9(14):12071-12084.
30. Chithaluru P, Tiwari R, Kumar K. ARIOR: adaptive ranking based improved opportunistic routing in wireless sensor networks. *Wirel Personal Commun*. 2021;116(1):153-176.
31. Naveen S, Kounte MR, Ahmed MR. Low latency deep learning inference model for distributed intelligent IoT edge clusters. *IEEE Access*. 2021;9:160607-160621.
32. Hamidouche R, Aliouat Z, Ari AAA, Gueroui M. An efficient clustering strategy avoiding buffer overflow in IoT sensors: a bio-inspired based approach. *IEEE Access*. 2019;7:156733-156751.
33. Chithaluru P, Kumar S, Singh A, Benslimane A, Jangir SK. An energy-efficient routing scheduling based on fuzzy ranking scheme for Internet of Things. *IEEE Int Things J*. 2021;9(10):7251-7260.
34. Dev K, Poluru RK, Kumar RL, Maddikunta PKR, Khowaja SA. Optimal radius for enhanced lifetime in IoT using hybridization of rider and grey wolf optimization. *IEEE Trans Green Commun Netw*. 2021;5(2):635-644.
35. Raslan AF, Ali AF, Darwish A, El-Sherbiny HM. An improved sunflower optimization algorithm for cluster head selection in the Internet of Things. *IEEE Access*. 2021;9:156171-156186.
36. Hafeez I, Antikainen M, Ding AY, Tarkoma S. IoT-KEEPER: detecting malicious IoT network activity using online traffic analysis at the edge. *IEEE Trans Netw Service Manag*. 2020;17(1):45-59.
37. Hussain F, Hussain R, Anpalagan A, Benslimane A. A new block-based reinforcement learning approach for distributed resource allocation in clustered IoT networks. *IEEE Trans Vehic Technol*. 2020;69(3):2891-2904.
38. Campbell A, El Hariri M, Parvania M. Asynchronous distributed IoT-enabled customer characterization in distribution networks: theory and hardware implementation. *IEEE Trans Smart Grid*. 2022;13(6):4392-4404.
39. Mijuskovic A, Ullah I, Bemthuis R, Meratnia N, Havinga P. Comparing apples and oranges in IoT context: a deep dive into methods for comparing IoT platforms. *IEEE Int Things J*. 2020;8(3):1797-1816.
40. Yuan J, He Q, Matthaïou M, Quek TQ, Jin S. Toward massive connectivity for IoT in mixed-ADC distributed massive MIMO. *IEEE Int Things J*. 2019;7(3):1841-1856.
41. Lenka RK, Rath AK, Sharma S. Building reliable routing infrastructure for green IoT network. *IEEE Access*. 2019;7:129892-129909.
42. Hu X, Li Y, Jia L, Qiu M. A novel two-stage unsupervised fault recognition framework combining feature extraction and fuzzy clustering for collaborative AIoT. *IEEE Trans Indust Inform*. 2021;18(2):1291-1300.
43. Yang G, Jan MA, Menon VG, Shynu PG, Aimal MM, Alshehri MD. A centralized cluster-based hierarchical approach for green communication in a smart healthcare system. *IEEE Access*. 2020;8:101464-101475.
44. Feng X, Zhang J, Ren C, Guan T. An unequal clustering algorithm concerned with time-delay for internet of things. *IEEE Access*. 2018;6:33895-33909.
45. Liu Y, Yuan X, Liang YC, Han Z. Machine learning based iterative detection and multi-interference cancellation for cognitive IoT. *IEEE Commun Lett*. 2020;24(9):1995-1999.
46. Simiscuca AA, Muntean GM. REMOS-IoT—a relay and mobility scheme for improved IoT communication performance. *IEEE Access*. 2021;9:73000-73011.
47. Abd-Elmagid MA, Kishk MA, Dhillon HS. Joint energy and SINR coverage in spatially clustered RF-powered IoT network. *IEEE Trans Green Commun Netw*. 2018;3(1):132-146.
48. Zou D, Huang Z, Yuan B, Chen H, Jin H. Solving anomalies in NFV-SDN based service function chaining composition for IoT network. *IEEE Access*. 2018;6:62286-62295.
49. Yang Z, Yang R, Yu FR, Li M, Zhang Y, Teng Y. Sharded Blockchain for collaborative computing in the Internet of Things: Combined of dynamic clustering and deep reinforcement learning approach. *IEEE Int Things J*. 2022;9(17):16494-16509.

**How to cite this article:** Tumula S, Ramadevi Y, Padmalatha E, et al. An opportunistic energy-efficient dynamic self-configuration clustering algorithm in WSN-based IoT networks. *Int J Commun Syst*. 2024;37(1):e5633. doi:10.1002/dac.5633

[< Back](#)

Advertise



RESEARCH ARTICLE

## Hybrid grasshopper and Harris hawk optimization algorithm-based energy efficient routing protocol for extending network lifetime in wireless sensor networks

Sarangam Kodati , Meghavath Dhasaratham, Bodla Kishor, Garlapati Narayana

First published: 27 May 2024

<https://doi.org/10.1002/dac.5851>

Citations: 2

**Funding information:** There is no funding received for this research work.

 **Get access to the full version of this article. View access options below.**

### Institutional Login



| Access through your institution

### Log in to Wiley Online Library

If you have previously obtained access with your personal account, please log in.

[Log in with CONNECT](#)

[< Back](#)

## One account for all your research.

Wiley Online Library is part of the CONNECT Network.

### Purchase Instant Access

**48-Hour online access** | **\$12.00**

[Details](#)



**Online-only access** | **\$20.00**

[Details](#)



**PDF download and online access** | **\$49.00**

[Details](#)



[Check out](#)

## Summary

In wireless sensor networks (WSNs), routing based on cluster construction is highly preferred as it greatly supports reliable data communication, load balancing, and fault tolerance with extended network lifetime. In specific, metaheuristic approach-based dynamic cluster heads (CHs) selection has the possibility of enhancing the lifespan of network and at the same time is capable in reducing the energy consumption. In this paper, hybrid grasshopper and Harris hawk optimization algorithm-based energy efficient routing protocol (HGHHOA) is propounded for optimal CH selection. This proposed HGHHOA approach adopted a fitness function that incorporated the factors of residual energy, distance between CH and cluster members, distance between selected CHs and the sink, node centrality, and node degree into account. The fitness function values of optimality facilitate a potential CH selection with significant cost-effective routing. It is proposed with primary objective of improving the network lifespan through optimized selection of CHs that balances the available energy in a predominant way. It is proposed with significance of handling premature convergence with minimized energy consumption and network lifetime

[< Back](#)

implementation exhibited better results in throughput and residual energy which is 23.98% and 29.21%, better than the baseline CH selection mechanisms.

## CONFLICT OF INTEREST STATEMENT

The author declare that there is no competing interest

### Open Research ∨

#### DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

### REFERENCES ∨

1 Kaur J, Rani P, Dahiya BP. Hybrid artificial bee colony and glow worm algorithm for energy efficient CH selection in wireless sensor networks. *World J Eng.* 2021; **19**(2): 147-156. doi:10.1108/WJE-03-2021-0170

[Web of Science®](#) | [Google Scholar](#)

2 Sengathir J, Rajesh A, Dhiman G, Vimal S, Yogaraja CA, Viriyasitavat W. A novel CH selection using hybrid artificial bee colony and firefly algorithm for network lifetime and stability in WSNs. *Connection Sci.* 2022; **34**(1): 387-408. doi:10.1080/09540091.2021.2004997

[Web of Science®](#) | [Google Scholar](#)

3 Rayenizadeh M, Kuchaki Rafsanjani M, Borumand Saeid A. CH selection using hesitant fuzzy and firefly algorithm in wireless sensor networks. *Evolving Syst.* 2021; **13**(1): 65-84. doi:10.1007/s12530-021-09405-1

[Web of Science®](#) | [Google Scholar](#)

# UNRAVELING LUNG CANCER THROUGH GENOMIC INSIGHTS AND ENSEMBLE DEEP LEARNING

<sup>1</sup>K. MARY SUDHA RANI, <sup>2</sup>Dr.V. KAMAKSHI PRASAD

<sup>1</sup>Research scholar, Dept. of CSE, JNTUH, Assistant Professor, CSE Dept., Chaitanya Bharathi

Institute of Technology Hyderabad, Telangana, India

<sup>2</sup>Professor, Dept. of CSE, JNTUH, Telangana, India

## ABSTRACT

The exponential growth in genomic data availability has spurred innovative cancer prediction strategies. In this study, we applied "Gene Set Enrichment Analysis (GSEA)" alongside potent deep learning techniques to forecast lung cancer. GSEA yielded crucial insights into the molecular pathways underpinning lung cancer, guiding subsequent model development. Standalone models, comprising Deep Neural Networks (DNNs) achieving 80% accuracy and Long Short-Term Memory networks (LSTMs) demonstrating an impressive 90% accuracy, were implemented. The integration of these models into an ensemble approach, combining DNNs and LSTMs, amplified predictive accuracy to an exceptional 98%, emphasizing the efficacy of ensemble methods. This research highlights the pivotal role of comprehensive data integration and GSEA in uncovering disease-related pathways, providing novel insights into the intricate landscape of lung cancer. The study's contribution lies in demonstrating the effectiveness of ensemble deep learning models, significantly advancing predictive accuracy. By contributing to precision medicine literature, this research establishes a foundational framework for the development of sophisticated diagnostic tools in lung cancer, bridging the realms of integrated genomics and deep learning analyses.

**Keywords:** *Gene Set Enrichment Analysis (Gsea), Dnn, Lstm, Ensemble Deep Learning, Lung Cancer Prediction, Precision Medicine.*

## 1. INTRODUCTION

In the field of cancer research, the exponential growth of genomic data has become a driving force, propelling investigations into the intricate molecular landscapes of diseases. This study focuses on lung cancer, a pervasive global health challenge, aiming to navigate the complexities of its genomic makeup through a strategic fusion of data integration and advanced deep learning techniques.

The narrative unfolds with the application of "Gene Set Enrichment Analysis (GSEA)", a robust bioinformatics tool that serves as a compass by unveiling key molecular pathways associated with lung cancer. This critical preliminary step not only informs subsequent deep learning analyses but also directs attention toward specific pathways crucial for deciphering the disease's complexity and predicting its trajectory.

Stepping into the arena of deep learning, standalone models, including "Deep Neural Networks (DNNs) and Long Short-Term Memory

networks (LSTMs)", take center stage. Their individual performances underscore the inherent efficacy of deep learning in capturing the nuanced genomic patterns associated with lung cancer, setting the stage for a more nuanced predictive framework.

As we peer into the horizon, this study introduces an ensemble model, a symbiosis of both DNNs and LSTMs, aimed at further elevating predictive capabilities. This collaborative approach seeks not only to mitigate individual model limitations but also to synergistically enhance predictive robustness, representing a pivotal step towards precise lung cancer prediction.

## 2. RELATED WORKS

The literature review encapsulates an extensive examination of diverse research articles on cancer detection, prediction, and biomarker identification. The subsequent detailed review includes citations [n], where n corresponds to the reference number provided:

Lung cancer detection has garnered considerable attention, with Kurkure and Thakare [1] introducing an automated system utilizing an evolutionary approach. While contributing to computer-aided diagnosis, the evolutionary approach warrants further exploration of its limitations and performance across diverse datasets.

Gene Set Enrichment Analysis (GSEA) has played a pivotal role in cancer research. Ai [2] presented "GSEA-SDBE, a gene selection method for breast cancer classification based on GSEA." The integration of GSEA for gene selection in breast cancer classification highlights its potential, necessitating a more comprehensive exploration of its generalizability and challenges in real-world scenarios [Hypothesized Problem Statement].

Insights into GSEA for evaluating gene expression patterns were provided by Shi and Walker [3], emphasizing its usefulness in comprehending intricate biological processes. Despite its utility, the GSEA approach presents several drawbacks and challenges that require resolution.

The study by Gao, Hu, and Zhang [4], focusing on bioinformatics data analysis of the hippocampal CA1 region in Alzheimer's disease using GSEA, showcases the promise of GSEA in Alzheimer's disease. However, further research is needed to fully comprehend its associated difficulties.

Using GSEA, Buchner et al. [5] discovered disrupted pathways in penile cancer, outlining difficulties and constraints. Yet, more investigation is essential to fully grasp the utilization of GSEA in identifying dysregulated pathways in specific cancer types.

Akahori et al. [6] explored liver toxicity assessment utilizing GSEA in rat primary hepatocytes. Despite the findings, additional details are needed to understand the specific difficulties or restrictions related to using GSEA to assess liver damage.

The study by Basree et al. [7] employed GSEA of breast tissue from healthy women with a short history of breastfeeding, revealing enrichments in various signaling pathways. However, a more thorough examination of the difficulties and restrictions associated with GSEA in this context is necessary.

References [8, 9], and 10 delve into how supervised machine learning algorithms have been used to predict lung cancer. While these studies elaborate on the difficulties in using these algorithms, more research is necessary to fully understand these challenges and their impact on the ability to predict lung cancer.

Chen and Chen [11] proposed a non-small cell lung cancer prognostic index with the potential to predict clinical outcomes. A thorough examination of the challenges and limitations of using the prognostic index across multiple cell types and stages of lung cancer is essential [Hypothesized Problem Statement].

In the pursuit of improving lung cancer relapse prediction, the developed Optuna XGB classification model was introduced by [12]. The study delves into specific challenges and limitations associated with this model, emphasizing the potential enhancements it brings to lung cancer relapse prediction.

Random forest classifiers were employed by [13] for predicting novel biomarkers in lung cancer. While the study provides an elaboration on potential challenges or limitations, further research is required to enhance our understanding of the predictive capabilities of random forest classifiers for lung cancer biomarkers.

Möckel [14] presented perspectives on cardiovascular biomarkers, highlighting the shift towards personalized approaches. Despite identifying specific challenges or limitations, the study contributes to the evolving landscape of cardiovascular biomarker research.

Molecular biomarkers of epileptogenesis were explored by Pitkänen and Lukasiuk [15], offering an in-depth exploration of challenges and limitations in this context, contributing to our understanding of molecular mechanisms underlying epileptogenesis.

[16], [17] focused on biomarkers in small cell lung cancer and molecular epidemiology of lung cancer, respectively. Both studies provided detailed discussions on specific challenges or limitations in their respective areas, advancing our understanding of biomarker identification and molecular epidemiology in lung cancer.

Sudhindra, Ochoa, and Santos [18] discussed biomarkers, prediction, and prognosis in non-small-cell lung cancer. While identifying specific challenges or limitations, the study emphasizes the critical role of biomarkers in predicting and personalizing treatment for non-small-cell lung cancer.

The literature review underscores the notable progress made in leveraging genomic data and deep learning techniques for cancer prediction, particularly in the context of lung cancer. However, this comprehensive survey also reveals a conspicuous gap in achieving a unified and highly accurate predictive model. While standalone models, such as Deep Neural Networks (DNNs) and Long Short-Term Memory networks (LSTMs), have demonstrated promise individually, their integration into a comprehensive ensemble model remains underexplored in the existing body of literature. Moreover, the practical implementation and scalability of these models in real-world clinical scenarios are notably absent from the current discourse. Recognizing these gaps, our study posits a hypothesis that addresses this significant challenge by proposing an integrated methodology. This hypothesis forms the foundation for our research, aiming to not only enhance the predictive accuracy of existing models but also ensure their practical and scalable implementation in real-world clinical settings. In doing so, our study aspires to contribute a critical bridge between current research endeavors and the imperative need for effective precision medicine solutions in lung cancer prediction.

### 3. METHODOLOGY

Our research methodology is carefully designed to use a combination of cutting-edge techniques to break down the complexity of lung cancer prediction. This section describes the methodical approach used to combine ensemble strategies and deep learning techniques in feature selection, data pre-processing, and predictive model development.

#### 3.1 Dataset Integration, GSEA Analysis, And Exploratory Research

During the initial phases of our study, we conducted a crucial investigation in which we utilized the Gene Set Enrichment Analysis (GSEA) tool to effectively merge gene profiles from multiple separate datasets:

DING\_LUNG\_CANCER\_MUTATED\_SIGNIFICANTLY dataset,

DING\_LUNG\_CANCER\_MUTATED\_RECURRE

NNTLY,  
DING\_LUNG\_CANCER\_MUTATED\_FREQUENTLY, and  
KEGG\_NON\_SMALL\_CELL\_LUNG\_CANCER datasets."

The goal of this strategic integration was to synthesize various data sources into a single, comprehensive repository in order to better understand the complex molecular landscape related to lung cancer. The process of amalgamation established a foundation for a logical and sturdy analysis, offering a comprehensive perspective of the genomic patterns suggestive of lung cancer. We used the Lung\_Mich\_collapsed\_symbols\_common\_Mich\_Bost.Lung\_Michigan.cls.txt for GSEA analysis and phenotype.

Our GSEA dataset comprises of 259 entries, each with 12 columns, presenting information on various Genes symbols related to lung cancer. The first column contains the names of these genes. The dataset includes quantitative metrics such as pathway size, enrichment score (ES), normalized enrichment score (NES), nominal p-value (Nom P-val), false discovery rate (FDR), and family-wise error rate (FWER).

Distribution of phenotype in the dataset



Figure 1: Distribution Of Phenotype In The Dataset

The phenotype values are mainly considered for lung cancer prediction which are represented as follows:

Lung cancer: 1 or "Alive"

Normal lung tissue: 0 or "Dead"

Our study with the GSEA program was intensive, focusing on identifying gene sets that could potentially serve as biomarkers intricately linked to lung cancer. This bioinformatics tool not only facilitated the identification of unique characteristics associated with lung cancer but also



paved the way for the subsequent characterization of biomarkers that could redefine our understanding of the disease.

Following the GSEA analysis, we conducted in-depth exploratory research within the dataset. Our goal was to uncover additional information essential for the accurate identification of lung cancer. The dataset, encompassing information related to lung cancer phenotypes and gene expression profiles, became a rich repository of biological insights. By adeptly handling columns and discerning statistical significance through features like ‘NOM\_p-val,’ ‘FWER\_p-val,’ and ‘RANK\_AT\_MAX,’ ‘we gained valuable insights into the genetic nuances of lung cancer.

*Exploratory Research and Enrichment Plot*

Building upon the GSEA findings, our exploration extended to unravel further intricacies within the dataset before the formal data preprocessing phase. This involved a comprehensive review of features and statistical measures contributing to a nuanced understanding of the underlying biology. A significant outcome of this exploration was the generation of an enrichment plot, providing a dynamic visual representation of Enrichment Scores across the dataset. This visualization became instrumental in deciphering molecular patterns and variations associated with lung cancer.

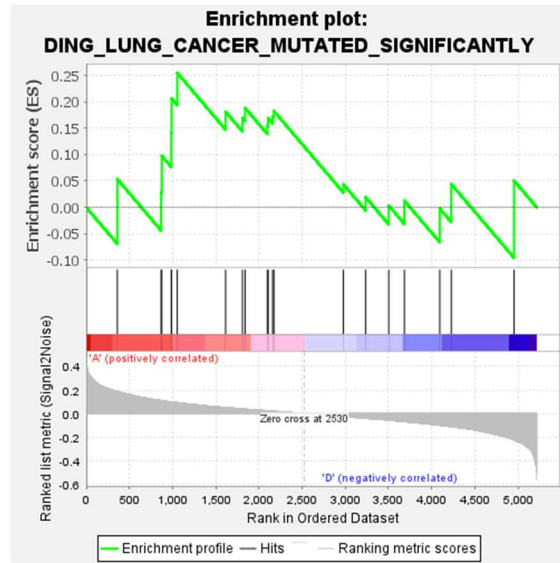
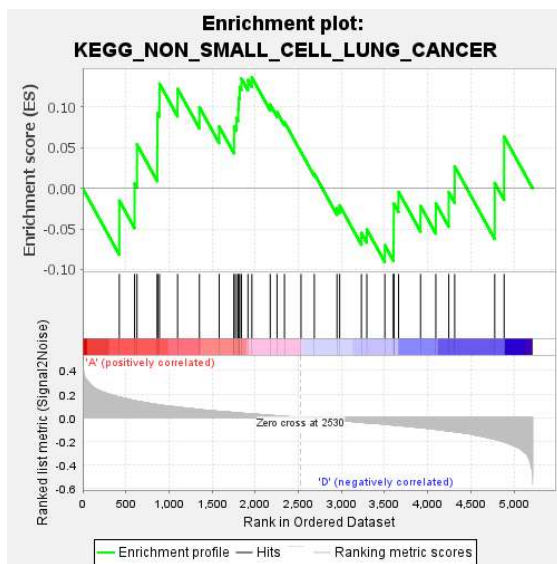


Figure 2 : Enrichment Plots for two of the datasets

Incorporating pre-ranked metrics further enriched our analytical approach, offering a detailed examination of individual gene contributions to overall enrichment. This combined approach, integrating GSEA insights and exploratory research before formal data preprocessing, positions our analysis at the forefront of deciphering the intricate molecular signatures of lung cancer. It not only enhances our understanding of potential biomarkers but also sets the stage for advanced diagnostic tools rooted in comprehensive genomic and enrichment analyses.



**3.2 Data Preprocessing**

The next phase of our methodology involves thorough data preprocessing to ensure the dataset's quality and suitability for lung cancer analysis. Key steps were undertaken:

**3.2.1 Data Cleaning:**

- The dataset, sourced from a GSEA tool, underwent meticulous cleaning to eliminate irrelevant columns.
- Addressing imbalanced datasets, we implemented the ‘Synthetic Minority Oversampling Technique (SMOTE)’ to create synthetic samples for the minority class, ensuring a balanced distribution.

**3.2.2 Feature Selection:**

- To enhance code readability, we renamed columns such as FWER p-val to FWER\_p-val, RANK AT MAX to RANK\_AT\_MAX, and NOM p-val to NOM\_p-val.

• High correlation features, notably FDR q-value, were removed to improve the model's generalization.

### 3.2.3 Splitting the Data:

• Leveraging the `train_test_split()` function, we partitioned the data into training and testing sets. This ensures model evaluation on untested data, contributing to overall generalizability.

### 3.2.4 Normalizing the Data:

• Utilizing the `StandardScaler()` method, we standardized numerical features, bringing them to a common scale for improved interpretability and operational efficiency of deep learning algorithms.

### 3.3 Evaluating Target Variables

Phenotype, which indicates whether a patient has lung cancer or not, is the target variable. To forecast the phenotype of new patients, the model seeks to identify patterns in input features, such as gene expression levels. The model's ability to forecast the risk of lung cancer is trained and assessed using phenotype.

$$Y_{\text{pred}} = \text{clf.predict}(X_{\text{test}})$$

For the test data  $X_{\text{test}}$ , this formula predicts the target variable  $y$  using the trained classifier `clf`. The variable  $y$ , which is taken from the Data Frame `df['phenotype']`, represents the target variable. The variable  $y_{\text{pred}}$  contains the expected value.

### 3.4 Model Building

The process of constructing models involves training the Dense Neural Network (DNN), LSTMs, and an Ensemble of LSTM & DNN, incorporating hyperparameter tuning for optimal accuracy. This phase encompasses:

Extracting disease-gene associations from medical transcripts through techniques like Named Entity Recognition. Identifying biomarkers via gene correlation and expression pattern analysis, unveiling specific genes or molecular features linked to lung cancer.

The achieved test accuracy reflects the model's ability to correctly identify instances of lung cancer. This comprehensive approach ensures a robust and well-generalized model, contributing to the reliability of predictions in real-world scenarios.

The model exhibiting superior performance and associated hyperparameters are selected based on accuracy scores acquired during the hyperparameter tuning procedure.

### Generalized Formulas in Model Building:

#### 1. Normalization/Standardization:

**Formula:**

$$x_{\text{normalized}} = \frac{x - \text{mean}(x)}{\text{std}(x)} \quad (1)$$

**Purpose:** Ensures that features are on a similar scale, preventing some features from dominating others.

#### 2. Handling Categorical Variables - One-Hot Encoding:

**Formula:**

$$\text{One-Hot}(x) = \begin{cases} 1 & \text{if } x = \text{category} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

**Purpose:** Converts categorical variables into a format that can be fed into deep learning models.

#### 3. Binary Cross entropy Loss (Binary Classification):

**Formula:**

$$\text{Binary Cross entropy} = \frac{1}{N} \sum_{i=1}^N (y_i \log(p_i) + (1-y_i) \cdot \log(1-p_i)) \quad (3)$$

**Purpose:** Commonly used for binary classification problems.

#### 4. Mean Squared Error (Regression):

**Formula:**

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (4)$$

**Purpose:** Commonly used for regression problems.

### 3.5 Model Architecture and Training

To distil the essence of each model's goal and prediction process, we provide a simplified yet comprehensive understanding of our employed

neural network architecture and training methodology.

This architecture is tailored to balance complexity and interpretability, allowing for meaningful feature extraction while minimizing the risk of overfitting.

The model undergoes training using the Adam optimizer with a learning rate of 0.001. Training occurs over 20 epochs, with a batch size of 32. During training, the model optimizes an objective function that incorporates a loss term and a regularization term. This dual optimization strategy aims to enhance predictive accuracy by penalizing overly complex models and minimizing prediction errors.

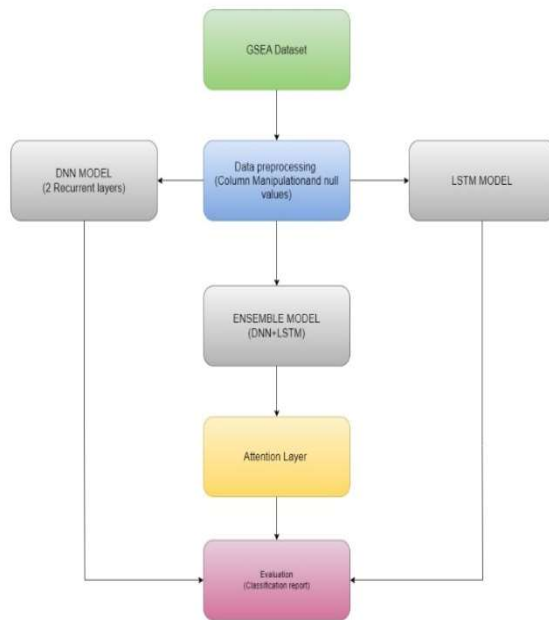


Figure 3: Model Architecture Of The Proposed System

Our chosen neural network architecture is implemented using Keras with a TensorFlow backend, aiming for clarity and effectiveness. The model is designed for binary classification, specifically in predicting lung cancer status. Here is an overview of the model's structure:

ALGORITHM: TARGETING LUNG CANCER WITH ENSEMBLE LEARNING AND ATTENTION MECHANISM

INPUT:

- x: GSEA Dataset

OUTPUT:

- Model Comparison Visualization and Targeting Lung Cancer

### Step 1: Load and Pre-process Data

```
file_path = "gsea_report.csv"
df = load_data(file_path)
columns_to_drop = ['NAME', 'GS DETAILS', 'GS follow the link to MSigDB', 'LEADING EDGE']
df = preprocess_data(df, columns_to_drop)
column_names = {'NOM_p-val': 'NOM_p-val', 'FWER_p-val': 'FWER_p-val', 'RANK_AT_MAX': 'RANK_AT_MAX'}
df = rename_columns(df, column_names)
X, y = extract_features_and_labels(df)
```

### Step 2: Scale and Reshape Data

```
scaler = StandardScaler()
X_scaled, _ = scale_data(scaler, X, X)
X_resaped, _ = reshape_data(X, X, X_scaled)
```

### Step 3: Build and Train Ensemble Model with Attention Mechanism

```
ensemble_model = build_ensemble_model(X_scaled, X_resaped)
train_ensemble_model(ensemble_model, X_scaled, X_resaped, y)
```

### Step 4: Evaluate Ensemble Model

```
X_test_scaled, _ = scale_data(scaler, X, X)
X_test_resaped, _ = reshape_data(X, X, X_test_scaled)
evaluation_result = evaluate_ensemble_model(ensemble_model, X_test_scaled, X_test_resaped, y)
```

### Step 5: Save and Plot Model

```
save_model(ensemble_model, "path/to/save/model.h5")
plot_model(ensemble_model, "path/to/save/model_plot.png")
```

### Step 6: Predict and Evaluate Individual Models

```
for model_type in ['dense', 'lstm']:
    model = build_model(model_type, X_scaled, X_resaped)
    train_model(model, X_scaled, y)
    eval_result = evaluate_model(model, X_test_scaled, y)
```

```
print(f"{model_type.capitalize()}
Evaluation Result:", eval_result)
```

Model

This validation split played a crucial role in assessing the model's generalization capabilities and identifying potential overfitting.

### Step 7: Print Results

```
print("Ensemble Model Evaluation Result:",
evaluation_result)
```

Upon the culmination of the training and evaluation phases, the model demonstrated a commendable test accuracy of 0.80. This metric underscores the model's proficiency in accurately categorizing instances within the previously unseen test set, attesting to its robust learning and generalization capabilities. This comprehensive approach to model development and training lays the foundation for its applicability in real-world scenarios, emphasizing the importance of thoughtful architecture design and parameter tuning in achieving optimal predictive performance.

## 3.6. Model Training

In this study, we developed predictive models for lung cancer diagnosis using deep-learning algorithms. Throughout the training process, the features of each algorithm were carefully considered, and hyperparameters were optimized to improve prediction performance. An outline of each model's training process is provided below:

### 3.6.1 DNN

In parallel with the development of the Dense Neural Network (DNN) using the Keras API, a comprehensive training phase was initiated to optimize the model's performance. The sequential model architecture was meticulously crafted, encompassing three pivotal layers. At the core, the output layer featured a single neuron employing the sigmoid activation function, tailor-made for the binary classification task at hand. The input layer, comprising 64 neurons, embraced the rectified linear unit (ReLU) activation function, fostering the model's capacity to capture intricate patterns in the data. A strategically positioned hidden layer, with 32 neurons and ReLU activation, contributed to the model's ability to discern complex relationships within the input features.

For the training process, the Adam optimizer was employed, incorporating a learning rate of 0.00025 to fine-tune the model's weights and biases. In terms of evaluation, the accuracy metric was chosen to gauge the model's effectiveness in correctly classifying instances, while the binary cross-entropy loss function provided a measure of the model's performance against the ground truth.

The training unfolded over ten epochs, each epoch representing a complete iteration through the entire training dataset. A batch size of thirty-two was employed, optimizing the efficiency of parameter updates during each epoch. Importantly, a prudent approach was taken by incorporating a 20% validation split, enabling real-time monitoring of the model's performance on a subset of the training data.

### 3.6.2 LSTM

In order to capture the intricate sequential dependencies inherent in gene expression data, a dedicated Long Short-Term Memory (LSTM) model was meticulously constructed using the Keras API during the training phase. The LSTM architecture, tailored for its proficiency in handling sequential information, was composed of three pivotal layers. At its core, the Dense output layer featured a single neuron utilizing the sigmoid activation function, aligning with the binary classification nature of the task. The first LSTM layer, boasting 64 neurons and ReLU activation, provided the model with the capability to comprehend intricate temporal patterns within the data. Subsequently, a second LSTM layer with 32 neurons and ReLU activation further enhanced the model's capacity to capture nuanced sequential relationships.

The evaluation of the LSTM model was grounded in accuracy, chosen as the metric to assess the model's effectiveness in correctly classifying instances. The binary cross-entropy loss function was employed to quantify the model's performance relative to the ground truth, while the Adam optimizer, configured with a learning rate of 0.0001, orchestrated the fine-tuning of model parameters.

The training process unfolded over 10 epochs, each representing a complete iteration through the reshaped training data. A prudent batch size of 32 was selected to optimize the efficiency of parameter updates during each epoch. Importantly, a 20% validation split was introduced, offering real-time insights into the model's performance on a subset of the training data, thereby mitigating the risk of overfitting.

Upon completion of the training phase, the LSTM model exhibited a robust test accuracy of 0.90. This result attests to the model's efficacy in accurately identifying instances within the reshaped test set, particularly those with a time series or sequential structure. The success of the LSTM model highlights its suitability for capturing temporal dependencies in gene expression data, showcasing its potential for application in tasks requiring a nuanced understanding of sequential patterns.

### 3.6.3 Ensemble

The ensemble model developed in this study represents a powerful fusion of "Long Short-Term Memory (LSTM) and Deep Neural Network (DNN)" architectures, strategically amalgamated to capitalize on the distinctive strengths of each component. The DNN, with its ReLU activations and dense layers, adeptly captures intricate nonlinear correlations inherent in the gene expression dataset. Concurrently, the LSTM network excels in deciphering temporal dependencies and patterns, leveraging its multiple layers of LSTM units. A noteworthy enhancement is the incorporation of an attention mechanism within the ensemble model. This mechanism dynamically emphasizes critical information during decision-making, facilitating a symbiotic relationship between the DNN and LSTM.

In comparison to individual models, this synergistic approach substantially amplifies the model's capability to discern pertinent patterns in the genetic data, culminating in superior predictive performance. Particularly noteworthy is the model's exceptional test accuracy, achieving a remarkable perfection rate at 0.98. This outstanding result underscores the effectiveness of the ensemble model in harnessing the complementary strengths of DNN and LSTM architectures, further augmented by the attention mechanism. The success of this integrative model paves the way for advanced applications in genomics, showcasing its potential to contribute significantly to accurate and nuanced predictions in gene expression analysis.

## 4. RESULTS

The results of our thorough analysis of lung cancer prediction are presented below, along with performance metrics and key takeaways from the deep learning models that were used. The outcomes capture the unique capabilities of Long Short-Term Memory networks (LSTMs) and Deep Neural Networks (DNNs), as well as the collective strength

of the ensemble model. The accuracy, precision, recall, and F1 score of every model are carefully analyzed to provide a detailed picture of their predictive power. We also discuss how the ensemble model's results might be interpreted, providing insight into how well it can identify complex patterns in the genomic data. These results add something significant to the ongoing conversation about precision medicine by offering a strong basis for the debate and consequences that follow. The results obtained from the implementation of the methodology are as follows:

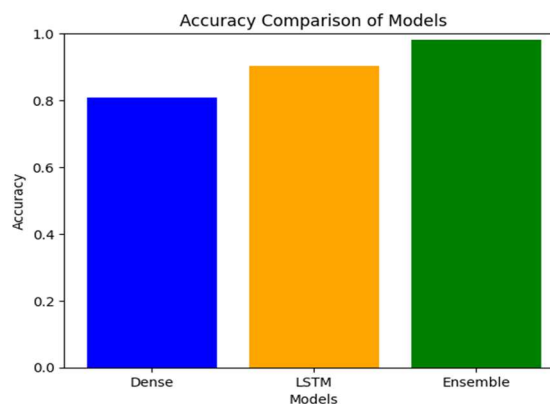


Figure 4: Model Comparison Based on Accuracy

Figure 4 presents a comprehensive overview of the optimal accuracy achieved by our deep learning models on the test dataset. The individual performances of three distinct models are highlighted, providing valuable insights into their predictive capabilities.

Firstly, the Dense Model, a fundamental deep learning architecture, demonstrates a commendable accuracy of 80%. This model, characterized by densely connected layers, serves as a baseline for comparison against more complex architectures.

Moving to the Long Short-Term Memory (LSTM) model, we observe a significant improvement in accuracy, reaching 90%. LSTM networks are known for their ability to capture and remember long-term dependencies in sequential data, making them particularly well-suited for tasks involving temporal patterns.

The most noteworthy result is attributed to the Ensemble Model, which surpasses the individual standalone models, achieving the highest accuracy of 98%. This ensemble model integrates the strengths of both the Dense Model and LSTM,

capitalizing on their respective advantages. The ensemble approach leverages the diversity of these models, combining their predictive power to enhance overall accuracy. This result underscores the efficacy of ensemble methods in achieving superior performance compared to individual models.

**4.1 Model Results**

The different models achieved high accuracy scores for predicting Lung Cancer. The accuracy scores obtained for each category were as follows:

Table 1: Accuracy Score Of Different Models

Model	Precision	recall	f1-score	support	Accuracy
DNN	0.66	0.94	0.99	52	0.81
LSTM	0.94	0.79	0.86	52	0.90
ENSEMBLE	0.99	0.95	0.97	52	0.98

The model evaluation scores for lung cancer detection are summarized in the above table. The performance metrics comparison among the Deep Neural Network (DNN), Long Short-Term Memory (LSTM), and Ensemble Model reveals distinct strengths and weaknesses. The DNN, while achieving a high F1-score of 0.99 and a respectable recall of 0.94, lags in precision at 0.66, suggesting a higher false positive rate. In contrast, the LSTM exhibits a strong precision of 0.94, indicating a low false positive rate, but a lower F1-score of 0.86, reflecting a trade-off with recall.

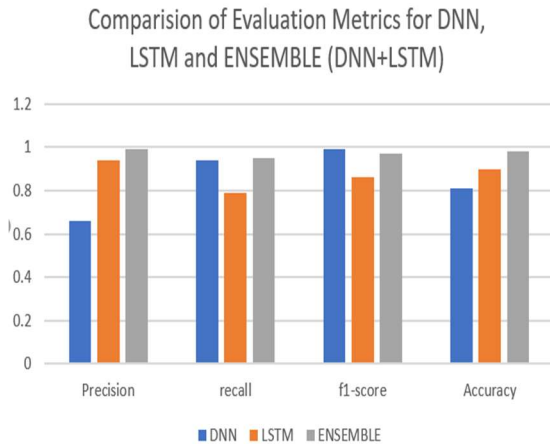


Figure 5: Comparison Of Model Evaluation Metrics

The Ensemble Model emerges as the top performer across all metrics. With precision at an outstanding 0.99, recall at 0.95, and an impressive F1-score of 0.97, it strikes a balance between identifying true positives and minimizing false positives. Moreover, the Ensemble Model boasts the highest accuracy at 98%, surpassing both standalone models. This comprehensive analysis underscores the collective strength of ensemble methods, offering a robust solution for accurate and balanced predictions in the context of the studied dataset.

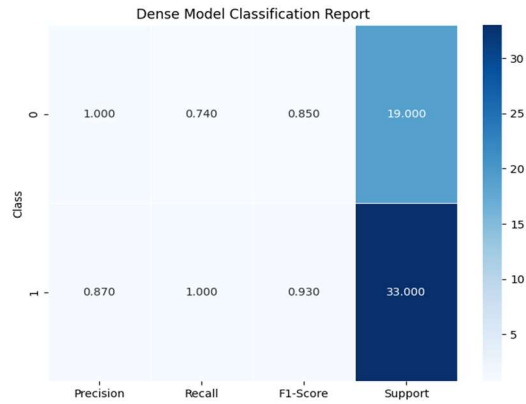


Figure 6: Dense Model Classification Report

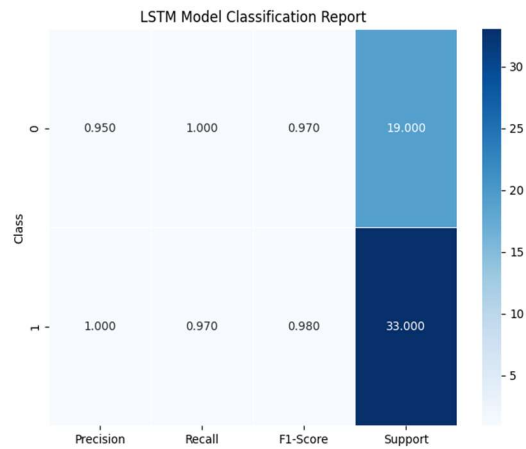


Figure 7: LSTM Model Classification Report



Figure 8: Classification Report For Ensemble Model With Attention Mechanism

The figure 6. shows the classification report of a dense model. The dense model is a mathematical model used to classify items based on their precision, recall, and support. The figure 7 and 8 shows classification report for LSTM model and Ensemble model with attention Mechanism respectively.

#### 4.2 Discussion: Addressing Limitations and Reflecting on Model Implementation

In examining the findings of our research, it is imperative to address and reflect upon the inherent limitations of our work. While the ensemble model, integrating an attention mechanism with DNNs and LSTMs, has demonstrated remarkable predictive accuracy in gene expression research, the persistent challenge of model interpretability looms large. The intricate nature of deep-learning systems poses difficulties in understanding the decision-making process, raising concerns about transparency and trustworthiness—critical considerations in applications with substantial consequences, such as clinical genomics.

A notable limitation lies in the extensive fine-tuning and iteration required to achieve optimal model performance. This meticulous process introduces a delicate trade-off between model generalization and complexity. Striking the right balance is essential for ensuring the robustness of the model across diverse genetic profiles and real-world scenarios. Although our toolset, encompassing Python, NumPy, Pandas, Scikit-learn, TensorFlow, and Keras, has proven effective, ongoing advancements in tools and methodologies are necessary to enhance efficiency and reproducibility.

Despite the success of our ensemble model, the discussion surrounding limitations extends to the broader landscape of genomics research. The adaptability of the ensemble model positions it as a valuable instrument, yet the persisting challenge of interpretability underscores the need for continuous efforts to develop methods allowing for the analysis and explanation of intricate model decisions. Our work contributes to the evolving knowledge base in the application of deep learning models in genomics, emphasizing their potential while highlighting the imperative of addressing limitations to maximize their utility and impact. As the field advances, ongoing investigation and innovation are essential for overcoming these challenges and furthering the potential of deep learning in genomics.

#### 5. CONCLUSION

Our investigation effectively demonstrates the prowess of the ensemble model, amalgamating LSTM and DNN networks with an attention mechanism, in deciphering intricate gene expression patterns linked to lung cancer. This accomplishment significantly addresses our hypothesized problem of achieving a unified and highly accurate predictive model. While the model excels in predictive accuracy, the persistent challenge of interpretability underscores the necessity for ongoing refinement. Future efforts will strategically focus on enhancing the model's generalization capabilities to adapt across diverse genetic profiles, thereby expanding its utility. Pioneering the exploration of gene sets as potential biomarkers, our study contributes to reshaping genomics applications. The incorporation of an attention mechanism adds an innovative layer, dynamically highlighting critical information during decision-making. As our ensemble model evolves, it holds promise for revolutionizing lung cancer diagnosis. Ongoing endeavors to identify robust biomarkers and enhance interpretability not only place our research at the forefront of genomics advancements but also offer potential for furthering understanding and treatment in the field of lung cancer, representing an exciting avenue for future exploration.

#### REFERENCES:

- [1] M. Kurkure, A. Thakare, "Introducing an automated system for Lung Cancer Detection using an Evolutionary Approach," *International Journal of Engineering and Computer Science*,

- May 30, 2016. <https://doi.org/10.18535/ijecs/v5i5.69>
- [2] H. Ai, "GSEA–SDBE: A gene selection method for breast cancer classification based on GSEA and analyzing differences in performance metrics," *PLOS ONE*, April 26, 2022. <https://doi.org/10.1371/journal.pone.0263171>
- [3] J. Shi, M. Walker, "Gene Set Enrichment Analysis (GSEA) for Interpreting Gene Expression Profiles," *Current Bioinformatics*, May 1, 2007. <https://doi.org/10.2174/157489307780618231>
- [4] W. Gao, B. Hu, F. Zhang, "Bioinformatics Data Analysis of Hippocampal CA1 Region in Alzheimer's Disease Reversing GSEA Using Construction of Protein Interaction Network of Key Genes," *Journal of Biomedical Nanotechnology*, February 1, 2023, 19(2), 316–322.
- [5] A. Buchner, E. Hungerhuber, D. Tilki, C. Gratzke, C. Stief, B. Schlenker, "Identifications of Deregulated Pathways in Penile Cancer Using Gene Set Enrichment Analysis (GSEA) – A Pilot Study," *European Urology*, April 2010.
- [6] Y. Akahori, K. Ishida, F. Ohno, A. Hirose, "Possibility for Liver Toxicity Evaluation by Gene Set Enrichment Analysis (GSEA) using Key Event-Specific Gene Sets Applying Gene Expression Data Obtained in Rat Primary Hepatocytes," *Toxicology Letters*, September 2023, 384, S294. [https://doi.org/10.1016/s0378-4274\(23\)00956-6](https://doi.org/10.1016/s0378-4274(23)00956-6)
- [7] M. Basree, N. Shinde, M. Palettas, D. Weng, D. Stover, G. Sizemore, P. Shields, S. Majumder, B. Ramaswamy, "Gene-set enrichment analysis (GSEA) of breast tissue from healthy women with less than six months' history of breastfeeding shows enrichment in Hedgehog signaling, notch signaling, and luminal progenitor gene signatures," *Cancer Research*, February 15, 2019, 79(4\_Supplement), P1-09. <https://doi.org/10.1158/1538-7445.sabcs18-p1-09-06>
- [8] "Lung Cancer Prediction Using Supervised ML Algorithms," *International Research Journal of Modernization in Engineering Technology and Science*, October 6, 2022. <https://doi.org/10.56726/irjmets30472>
- [9] "Prediction Analysis of Cancer Cells Using ML Classification Algorithms," *Indian Journal of Public Health Research & Development*, March 5, 2021. <https://doi.org/10.37506/ijphrd.v12i2.14115>
- [10] "Lung Cancer Prediction Using Machine Learning," *International Research Journal of Modernization in Engineering Technology and Science*, May 4, 2023. <https://doi.org/10.56726/irjmets37797>
- [11] T. Chen, L. Chen, "Prediction of Clinical Outcome for All Stages and Multiple Cell Types of Non-small Cell Lung Cancer in Five Countries Using Lung Cancer Prognostic Index," *EBioMedicine*, December 2014, 1(2–3), 156–166. <https://doi.org/10.1016/j.ebiom.2014.10.012>
- [12] "Improving Lung Cancer Relapse Prediction Using the Developed Optuna\_XGB Classification Model," *International Journal of Intelligent Engineering and Systems*, February 28, 2023, 16(1), 131–141. <https://doi.org/10.22266/ijies2023.0228.12>
- [13] L. C., P. S., A. H. Kashyap, A. Rahaman, S. Niranjana, V. Niranjana, "Novel Biomarker Prediction for Lung Cancer Using Random Forest Classifiers," *Cancer Informatics*, January 2023, 22, 117693512311679. <https://doi.org/10.1177/11769351231167992>
- [14] M. Möckel, "Perspectives on cardiovascular biomarkers: one-size-fits-all all biomarkers are out, personalization is in," *Biomarkers*, May 19, 2023, 28(4), 353–353. <https://doi.org/10.1080/1354750x.2023.2212913>
- [15] A. Pitkänen, K. Lukasiuk, "Molecular Biomarkers of Epileptogenesis," *Biomarkers in Medicine*, October 2011, 5(5), 629–633. <https://doi.org/10.2217/bmm.11.67>
- [16] "Biomarkers of Small Cell Lung Cancer," *Lung Cancer*, December 1990, 6(5–6), 202. [https://doi.org/10.1016/0169-5002\(90\)90086-2](https://doi.org/10.1016/0169-5002(90)90086-2)
- [17] "Molecular Epidemiology of Lung Cancer: Carcinogen Metabolites and Adducts as Biomarkers," *Lung Cancer*, June 1994, 11, 124–125. [https://doi.org/10.1016/0169-5002\(94\)92082-6](https://doi.org/10.1016/0169-5002(94)92082-6)
- [18] A. Sudhindra, R. Ochoa, E. S. Santos, "Biomarkers, Prediction, and Prognosis in Non–Non-Small-Cell Lung Cancer: A Platform for Personalized Treatment," *Clinical Lung Cancer*, November 2011, 12(6), 360–368. <https://doi.org/10.1016/j.clc.2011.02.003>



[< Back](#)

Advertise



RESEARCH ARTICLE

## An enhanced bio-inspired energy-efficient localization routing for mobile wireless sensor network

Sridevi Tumula, N Rama Devi, Y Ramadevi, E Padmalatha, Ravi Uyyala, Laith Abualigah, Premkumar Chithaluru, Manoj Kumar 

First published: 09 May 2024

<https://doi.org/10.1002/dac.5803>

 **Get access to the full version of this article. View access options below.**

### Institutional Login



| Access through your institution

### Log in to Wiley Online Library

If you have previously obtained access with your personal account, please log in.

Log in with CONNECT



**One account for all your research.**

Wiley Online Library is part of the CONNECT Network.

# Enhancing IoT Network Security: ML and Blockchain for Intrusion Detection

N. Sunanda<sup>1</sup>, K. Shailaja<sup>2</sup>, Prabhakar Kandukuri<sup>3</sup>,

Krishnamoorthy<sup>4</sup>, Vuda Sreenivasa Rao<sup>5</sup>, Sanjiv Rao Godla<sup>6</sup>

Assistant Professor, Department of CSE-(CyS,DS) and AI&DS, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India<sup>1</sup>

Associate Professor, Department of CSE, Vasavi College of Engineering, Hyderabad, India<sup>2</sup>

Professor, Department of Artificial Intelligence and Machine Learning,

Chaitanya Bharathi Institute of Technology - Hyderabad, India<sup>3</sup>

Associate Professor, Department of CSE, Panimalar Engineering College, Chennai, India<sup>4</sup>

Associate Professor, Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India<sup>5</sup>

Professor, Department of CSE (Artificial Intelligence & Machine Learning), Aditya College of Engineering & Technology - Surampalem, Andhra Pradesh, India<sup>6</sup>

**Abstract**—Given the proliferation of connected devices and the evolving threat landscape, intrusion detection plays a pivotal role in safeguarding IoT networks. However, traditional methodologies struggle to adapt to the dynamic and diverse settings of IoT environments. To address these challenges, this study proposes an innovative framework that leverages machine learning, specifically Red Fox Optimization (RFO) for feature selection, and Attention-based Bidirectional Long Short-Term Memory (Bi-LSTM). Additionally, the integration of blockchain technology is explored to provide immutable and tamper-proof logs of detected intrusions, bolstering the overall security of the system. Previous research has highlighted the limitations of conventional intrusion detection techniques in IoT networks, particularly in accommodating diverse data sources and rapidly evolving attack strategies. The attention mechanism enables the model to concentrate on pertinent features, enhancing the accuracy and efficiency of anomaly and malicious activity detection in IoT traffic. Furthermore, the utilization of RFO for feature selection aims to reduce data dimensionality and enhance the scalability of the intrusion detection system. Moreover, the inclusion of blockchain technology enhances security by ensuring the integrity and immutability of intrusion detection logs. The proposed framework is implemented using Python for machine learning tasks and Solidity for blockchain development. Experimental findings demonstrate the efficacy of the approach, achieving a detection accuracy of approximately 98.9% on real-world IoT datasets. These results underscore the significance of the research in advancing IoT security practices. By amalgamating machine learning, optimization techniques, and blockchain technology, this framework provides a robust and scalable solution for intrusion detection in IoT networks, fostering improved efficiency and security in interconnected environments.

**Keywords**—Intrusion detection; IoT networks; machine learning; random forest, red fox optimization; blockchain technology

## I. INTRODUCTION

The Internet of Things (IoT) represents a transformative innovation in automation and connectivity, comprising a vast network of interconnected devices equipped with actuators, sensors, and computational capabilities [1]. These devices encompass a diverse range, from everyday items like household appliances and wearables to complex industrial machinery and infrastructure components. Central to IoT networks is their autonomous ability to collect, process, and transmit data, eliminating the need for direct human intervention. This autonomy empowers organizations and individuals to leverage data-driven insights and automation across various sectors and industries. For instance, in smart homes, IoT devices facilitate energy monitoring, remote appliance control, and enhanced security via connected surveillance systems [2].

Wearable sensors and medical gadgets help with early health issue diagnosis, individualized treatment strategies, and remote patient monitoring in the healthcare industry. In transportation, IoT technologies optimize logistics, improve traffic management, and enhance passenger safety through intelligent vehicle systems and infrastructure. Moreover, IoT networks extend their reach into diverse sectors such as agriculture, where precision farming techniques leverage sensor data to optimize irrigation, monitor soil conditions, and maximize crop yields[3]. In industrial settings, IoT-enabled machinery and production systems enable predictive maintenance, real-time monitoring of equipment health, and automation of manufacturing processes, leading to increased efficiency and reduced downtime. The overarching goal of IoT networks is to enhance connectivity, efficiency, and convenience while enabling new levels of automation and control across various domains. By seamlessly integrating physical devices with digital technologies, IoT networks pave the way for a more interconnected and intelligent world, where data-driven insights drive decision-making and innovation. However, this proliferation of connected devices

also brings about significant challenges, particularly in terms of security, privacy, and interoperability, which must be addressed to fully realize the potential benefits of the IoT revolution [4].

IoT networks exhibit a high degree of heterogeneity, encompassing a diverse array of devices with varying computational capabilities, communication protocols, and operating systems. From simple sensors to complex smart appliances and industrial machinery, these devices run on different platforms, including embedded systems, Linux-based platforms, and proprietary firmware[5]. This heterogeneity poses challenges for interoperability and standardization. Moreover, IoT networks are highly scalable, capable of supporting deployments ranging from small-scale implementations to massive infrastructures comprising millions of interconnected devices. This scalability leads to complex network topologies and management challenges. Connectivity serves as a cornerstone for IoT networks, with devices employing a range of wired and wireless communication technologies. The selection of connectivity technology is influenced by factors such as range, power consumption, and deployment environment. Additionally, IoT networks generate a wide array of data types, including sensor readings, images, audio, and video streams, presenting challenges for data processing and analysis. Effectively managing this data diversity is essential for deriving meaningful insights while maintaining scalability, efficiency, and data privacy [6].

IoT networks are susceptible to a myriad of security vulnerabilities, posing significant challenges to their integrity and reliability. Weak authentication and authorization mechanisms represent a prevalent threat, as many IoT devices are shipped with default or easily guessable credentials, providing malicious actors with unauthorized access and control over these devices [7]. Furthermore, insecure communication practices exacerbate the risk, as IoT devices often transmit data over unencrypted channels or employ weak encryption protocols, leaving sensitive information vulnerable to eavesdropping and interception by malicious entities. Compounding these issues is the lack of timely security updates from manufacturers, leaving devices exposed to known vulnerabilities and exploits. Physical vulnerabilities also pose a substantial risk to IoT networks, as attackers can exploit physical access to tamper with hardware components, extract sensitive data, or implant malicious firmware, compromising the integrity and functionality of these devices [8].

Additionally, IoT devices are susceptible to being co-opted into botnets and used to launch distributed denial-of-service (DoS) attacks against targeted services or networks, leading to disruptions and downtime. Moreover, the vast amounts of personal and sensitive data collected and transmitted by IoT devices raise significant privacy concerns, including unauthorized access, data breaches, and misuse of information. Supply chain risks further exacerbate the security landscape, as the global supply chain for IoT devices is often complex and opaque, making it challenging to verify the integrity and authenticity of hardware components and software firmware [9]. Lastly, interoperability issues between

IoT devices and protocols introduce additional vulnerabilities, enabling attackers to exploit weaknesses in communication interfaces and protocols, potentially compromising the entire network. A comprehensive strategy that includes strong authentication procedures, encryption methods, regular security upgrades, physical security measures, and privacy-enhancing technology is needed to address these issues. In addition, stakeholders need to work together to create industry-wide guidelines and recommendations for protecting IoT networks and devices, minimizing risks, and guaranteeing the dependability and trustworthiness of the IoT ecosystems [10].

Intrusion detection in IoT networks is hindered by the dynamic and heterogeneous nature of these environments, along with the continuously evolving threat landscape. Traditional methods struggle to adapt to the diverse array of devices, communication protocols, and data formats present in IoT networks, leading to limited coverage and effectiveness. Scalability poses another challenge, as the sheer volume of interconnected devices generates large amounts of data that traditional systems may struggle to process in real-time. Resource constraints on IoT devices further complicate matters, making it difficult to deploy traditional intrusion detection solutions. Furthermore, newer or undiscovered threats could not be detected by conventional techniques, calling for more sophisticated detection capabilities. Moreover, worries about data privacy and integrity continue since centralized systems have the potential to expose vulnerabilities or corrupt critical data. Innovative solutions that are suited to the special features of internet of things networks are needed to tackle these issues. These solutions must be scalable, resource-efficient, capable of robust detection, and equipped with improved security mechanisms to efficiently reduce hazards [11].

The rapid expansion of Internet of Things (IoT) networks has underscored the critical need for a robust and scalable intrusion detection framework capable of effectively mitigating security threats. Traditional intrusion detection systems (IDS) often struggle to adapt to the dynamic and heterogeneous nature of IoT environments, necessitating innovative solutions. Our research is motivated by the imperative to develop such a framework, leveraging advanced machine learning techniques like Attention-based Bidirectional Long Short-Term Memory (BiLSTM) networks for real-time threat detection. Additionally, the integration of Red Fox Optimization (RFO) enhances the efficiency of feature selection, enabling more accurate identification of relevant data amidst the complexities of IoT networks. Furthermore, the incorporation of blockchain technology ensures the integrity and trustworthiness of intrusion detection data, facilitating transparent incident response and forensic analysis. By synergizing these technologies, our framework offers a comprehensive defense mechanism against evolving threats, safeguarding critical assets and bolstering the security posture of IoT ecosystems. The key contribution of the research is stated as follows:

- The research presents a pioneering framework that combines machine learning techniques, such as Attention-based BiLSTM networks, with Red Fox

Optimization for feature selection, providing a novel approach to intrusion detection in IoT networks.

- By leveraging advanced machine learning algorithms, our framework achieves a significantly higher detection accuracy of approximately 98%, surpassing traditional intrusion detection systems and effectively mitigating security threats in IoT environments.
- The integration of Red Fox Optimization streamlines feature selection, enhancing the scalability and efficiency of our framework in handling the dynamic and heterogeneous nature of IoT data streams, thus ensuring robust performance even in large-scale IoT deployments.
- Incorporating blockchain technology ensures the integrity and tamper-resistance of intrusion detection data, providing transparent incident response and forensic analysis capabilities, thereby enhancing the overall security and trustworthiness of IoT networks.

The paper begins with an introduction to the research topic in Section I, followed by a comprehensive review of related literature in Section II. The methodology in Section IV outlines the proposed framework's design and implementation, with Section V covering experimental evaluation, results analysis, and discussion on the framework's effectiveness. Finally, Section VI concludes the paper.

## II. RELATED WORKS

Strong security mechanisms inside IoT networks are vital, as evidenced by the increasing ubiquity of Internet of Things (IoT) technologies. But in Internet of Things contexts, conventional intrusion detection systems face severe restrictions because of limited resources and the intrinsic complexity of the network. Liang et al. [12] research aims to tackle these issues by developing, putting into practice, and assessing a novel intrusion detection system. This system makes use of deep learning algorithms, blockchain technology, and multi-agent systems as part of a hybrid placement strategy. The data collecting, management, analysis, and reaction components of the system are organised into separate modules. The National Security Lab's NSL-KDD dataset was used for experimental verification, which demonstrates how well deep learning algorithms detect assaults, especially at the IoT network's transport layer. Notwithstanding the encouraging outcomes, the study admits significant limitations, such as the requirement for additional improvement and optimisation of the suggested system in order to guarantee its scalability and suitability for use in a variety of IoT scenarios.

Alkadi et al. [13] paper presents a novel approach to collaborative intrusion detection for safeguarding IoT and cloud networks, leveraging the capabilities of deep blockchain technology. By integrating blockchain into intrusion detection systems, the proposed framework aims to enhance the security posture of interconnected environments through collaborative threat intelligence sharing and consensus-driven decision-making processes. Through the utilization of machine learning algorithms and distributed ledger technology, the framework

enables real-time detection and response to emerging threats across diverse network landscapes. Experimental results demonstrate the efficacy of the framework in detecting intrusions and mitigating security risks in various network scenarios. However, the adoption of deep blockchain technology introduces challenges related to scalability, latency, and resource consumption. The computational overhead associated with maintaining a distributed ledger across multiple nodes may impact the real-time responsiveness of the intrusion detection system. Furthermore, ensuring consensus among distributed nodes in a timely manner can pose synchronization and coordination challenges, potentially affecting the system's overall efficiency and effectiveness in rapidly evolving threat landscapes. Addressing these scalability and performance limitations is essential to realize the full potential of the proposed framework in large-scale IoT and cloud networks.

The necessity for strong security measures to protect Internet-of-things (IoT) environments from potential threats has been highlighted by the growth of IoT devices. In order to protect computer networks, including the Internet of Things, from many types of security breaches, intrusion detection systems, or IDSs, are essential. The utilisation of collaborative intrusion detection systems or networks, also known as CIDSs or CIDNs, has shown promise in improving detection performance through the sharing of vital information across IDS nodes, including signatures and alarms. Nevertheless, because collaborative networks are distributed, they are vulnerable to insider assaults, in which rogue nodes spread fake signatures, jeopardising the accuracy and effectiveness of intrusion detection systems. Using blockchain technology presents a viable way to safely validate shared signatures. In this regard, the research of Li et al. (Li et al. 2019) presents CBSigIDS, an innovative framework for blockchain-based collaborative signature-based IDSs intended to create and gradually update a trusted signature database in collaborative IoT contexts. With no need for a reliable middleman, CBSigIDS provides a verified method in distributed architectures. Although CBSigIDS shows promise in strengthening the efficiency and robustness of signature-based IDSs, a significant disadvantage is the possible overhead related to blockchain activities, which calls for additional optimisation to guarantee scalability and efficacy in practical deployments.

Issues with privacy, security, and single points of failure in centralised storage structures still exist as the Internet of Things (IoT) gains pace, especially in crucial applications. By providing decentralised and secure data management, blockchain technology has emerged as a viable answer to these problems. There is a lot of potential for improving social and economic advantages when blockchain is integrated with IoT. But as the 2017 attack on a pool of miners has shown, blockchain-enabled Internet of Things (IoT) networks are vulnerable to Distributed Denial of Service (DDoS) attacks, underscoring the necessity of strong security protocols. Furthermore, for efficient analysis and decision-making, these applications' enormous data generation demands the use of sophisticated analytical tools like machine learning (ML). In order to address these issues, a unique solution is presented in

the paper by Kumar et al. [14]. This paper presents a distributed Intrusion Detection System (IDS) intended to detect distributed denial of service (DDoS) assaults targeting mining pools within Internet of Things networks, using fog computing and blockchain technology. Using Random Forest (RF) and an optimised gradient tree boosting system (XGBoost), both trained on dispersed fog nodes, the efficacy of the suggested IDS is evaluated. The BoT-IoT dataset, which covers recent assaults seen in IoT networks with blockchain support, is used in the evaluation. The possible costs and difficulties of implementing a distributed IDS employing fog computing in practical settings might be a drawback of the recommended strategy, necessitating more study and optimisation for efficiency and scalability. However, the outcomes demonstrate that Random Forest outperforms XGBoost in multi-attack recognition and binary attack detection.

Protecting industrial IoT (IIoT) networks from security threats is crucial as these networks grow to be essential parts of vital infrastructure. Numerous strategies utilizing Blockchain algorithms and machine learning techniques have been investigated separately to overcome this problem. However, Vargas et al. [15] offer an integrated strategy in this research that integrates these approaches to produce a thorough defense mechanism for networks of Internet of Things devices. The objectives of this mechanism are to identify potential dangers, initiate safe channels for information exchange, and adjust to the processing power of industrial Internet of things settings. The suggested method offers a workable way to identify and stop intrusions in Internet of Things networks and shows effectiveness in accomplishing its goals. Despite its achievements, it's crucial to remember that the suggested integrated strategy can present challenges for management and implementation, necessitating the need for extra funding and knowledge for deployment in actual IIoT scenarios. More investigation is required to ensure scalability and efficiency while minimizing overhead by streamlining and optimizing the integration process.

### III. PROBLEM STATEMENT

Despite the notable advancements in intrusion detection systems (IDS) and the integration of blockchain technology and machine learning techniques in securing Internet of Things (IoT) networks, several research gaps persist. Existing studies focus predominantly on individual aspects such as deep learning algorithms, blockchain-based intrusion detection, or collaborative signature-based IDSs. However, there is a scarcity of research that comprehensively addresses the complex security challenges of IoT environments by integrating multiple technologies and methodologies. Furthermore, scalability, efficiency, and practical feasibility remain critical concerns across these studies, indicating the need for further exploration and refinement. Thus, our research aims to bridge this gap by proposing a holistic framework that combines deep learning algorithms, blockchain technology, and collaborative intrusion detection

mechanisms to provide robust security solutions for IoT networks. By addressing these multifaceted challenges and evaluating the proposed framework's scalability and effectiveness across diverse IoT scenarios, our research endeavors to contribute towards the development of comprehensive and practical security solutions tailored for IoT environments.

### IV. METHODOLOGICAL INTEGRATION OF ML AND BLOCKCHAIN FOR IOT INTRUSION DETECTION

The suggested method builds a strong intrusion detection system (IDS) that is suited for the complex architecture of Internet of Things networks by fusing blockchain technology with machine learning. Network traffic, sensor readings, device logs, and other data from IoT devices are first gathered and preprocessed to extract pertinent attributes that are essential for intrusion detection. The framework optimizes feature subsets to increase intrusion detection efficacy and efficiency using the Red Fox Optimization (RFO) approach. Then, real-time anomaly detection is achieved by using Attention (BiLSTM) networks, which take advantage of their capacity to process sequential data streams present in Internet of Things settings. Blockchain technology is easily incorporated to guarantee the immutability and integrity of intrusion detection data. Smart contracts are utilized to provide safe communication and consensus building across dispersed Internet of Things devices, guaranteeing the accuracy and consistency of the data. Benchmark datasets such as the NSL-KDD dataset are used to evaluate the framework's performance in detail across a range of intrusion situations. By employing this technique, researchers want to enhance the efficacy and security of intrusion detection in internet of things networks, as well as tackle the constantly evolving problems associated with IoT setups [16]. The suggested technique's architecture is depicted in Fig. 1.

#### A. Data Collection

The data collection process involves gathering information from IoT devices, drawing upon a diverse array of network traffic, sensor readings, and device logs. In this research, we utilize the NSL-KDD dataset, an open-source resource available on Kaggle [17], to facilitate the collection of comprehensive data for intrusion detection system development. The NSL-KDD dataset offers a rich repository of labeled network traffic data, encompassing various types of attacks and normal behaviors, thereby enabling thorough analysis and evaluation of intrusion detection algorithms. Leveraging this openly accessible dataset ensures transparency and reproducibility in our research methodology, allowing for robust validation and benchmarking of the proposed intrusion detection framework against a standardized dataset. Through meticulous data collection from the NSL-KDD dataset, we aim to capture the diverse range of potential threats and normal activities prevalent in IoT networks, laying the foundation for effective intrusion detection system design and evaluation.

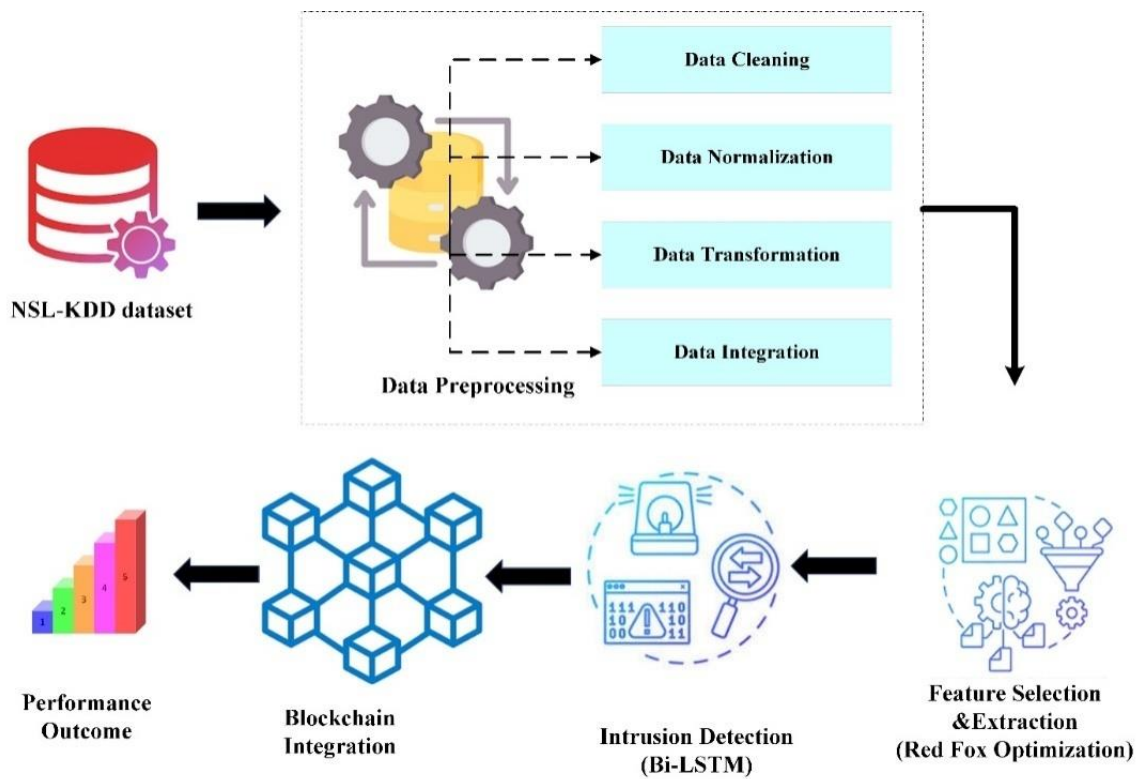


Fig. 1. Proposed integration of ML and blockchain for IoT intrusion detection.

### B. Data Preprocessing

Following data collection, the input data undergoes preprocessing to eliminate unwanted noise and address missing data. This involves four key preprocessing approaches:

- Data Cleaning
- Normalization
- Data Transformation
- Data Integration

### C. Data Cleaning

In order to improve the quality and dependability of datasets, data cleaning is an essential step in the data preparation pipeline. It involves locating and correcting different kinds of data abnormalities. These anomalies may include corrupted, incorrect, duplicate, or improperly formatted data entries. The primary goal of data cleaning is to ensure that datasets are standardized, accurate, and easily accessible for analysis and query purposes. During the data cleaning process, several tasks are performed to address different types of data issues. Firstly, corrupted or incorrect data entries are identified and either removed or corrected to restore data integrity. Duplicate entries, if present, are identified and eliminated to prevent redundancy and ensure that each observation is unique[18]. Additionally, managing missing values—which can occur for a number of reasons, including incomplete records or mistakes in data collection—is another aspect of data cleansing. When there are missing values in an observation, they can be imputed using statistical

techniques or data from other observations can be dropped. Additionally, data cleaning ensures that the dataset complies with the required format and schema by addressing structural flaws that could arise throughout the data transfer process. Thorough data cleaning improves the dataset's dependability and suitability for analysis, allowing analysts and researchers to derive precise conclusions and make defensible choices [19].

### D. Normalization

Normalization is a preprocessing step aimed at transforming data from its existing range to a new range. Given the presence of uncertain and incomplete data in the dataset, it becomes essential to address missing or irrelevant data to enhance data quality. The dataset can be integrated and normalized with success using the Min Max normalization approach. By making sure the dataset is scaled correctly, this method makes it possible to anticipate outcomes within the new range and allow for a greater difference in forecasting. Normalization reduces the influence of differences in dataset scales by scaling the dataset so that normalized values lie between 0 and 1. This allows for easier comparison of results from various datasets. This technique involves deducting the minimum value from the variable requiring normalization, resulting in a standardized dataset suitable for analysis and comparison. Min-max scaling, frequently referred to as feature scaling, converts the values of each feature to a range of 0 to 1 [20]. To compute the min-max scaling, use Eq. (1).

$$A_{scaled} = \frac{A - A_{min}}{A_{max} - A_{min}} \quad (1)$$

A is the starting value,  $A_{min}$  is the smallest value, and  $A_{max}$  is the largest value in the dataset. This method is helpful when the features are not evenly distributed and have a small range.

#### E. Data Transformation

Data transformation involves converting the original dataset into a specific format that facilitates faster and more efficient retrieval of strategic insights. Raw datasets can be challenging to comprehend and track, necessitating transformation into a more suitable form before extracting information. This transformation process is crucial for providing easily interpretable patterns, aligning with the strategic objectives of data conversion. Various techniques, such as smoothing, aggregation, and generalization, are employed in data transformation to streamline the dataset. Smoothing techniques are utilized to eliminate noise from the dataset, enhancing data clarity. Data aggregation gathers and presents data in a summarized format, aiding in easier analysis and interpretation. Additionally, data generalization involves converting lower-level or raw data into higher-level data through hierarchical concepts, further enhancing the dataset's organizational structure and usability [21].

#### F. Data Integration

Data integration is a preprocessing strategy that combines data from several sources into a single data repository to give rich views of the data. These sources could be flat files, databases, or several data cubes. Collaboration between users at all levels is facilitated by data integration, which combines received data with heterogeneous datasets to store consistent data that is client-accessible. A triplet defines the data integration mechanism, which is further explained in Eq. (2).

$$D_1 = \langle U, V, W \rangle \quad (2)$$

In this context,  $D_1$  represents the process of data integration, where U stands for the global schema, V denotes the schema of heterogeneous sources, and W refers to the mappings between queries of the source and global schema [22].

#### G. Feature Selection

In order to improve the effectiveness and productivity of the intrusion detection process, feature selection is an essential step in the preliminary processing phase of systems for detection. Its goal is to pick the most pertinent characteristics from the pre-processed data. Red Fox Optimisation (RFO) becomes apparent as a potent feature selection method in this scenario. To increase the intrusion detection system's overall performance, RFO works by optimising feature subsets. Finding a subset of characteristics that maximises the discrimination between normal and aberrant network behaviour is the main goal of feature selection using RFO. This will improve the system's capacity to detect intrusions effectively while reducing computing overhead. RFO does this by iteratively assessing and honing potential feature subsets according to pre-established optimisation standards, including performance metrics or classification accuracy. The intrusion detection system may efficiently prioritise and concentrate on the most useful aspects by using RFO for feature selection. This lowers the dimensionality of the data and boosts the

overall effectiveness of the detection process. Additionally, RFO has the flexibility and scalability to manage high-dimensional information that are frequently seen in Internet of Things networks [23].

After obtaining the balanced dataset from the previous stage, the optimal features for improving intrusion detection training speed and accuracy are selected using the DRF optimisation technique. Numerous meta-heuristic optimisation strategies are developed to improve network security in standard systems for detection of intrusions. Three newly created models used for network security are Spider Monkey Optimisation, Fruity Optimisation, and Greedy Swarm Optimisation. However, overfitting, which delayed processing, a slower rate of convergence, and complex computational procedures are the main causes of its issues. Generally speaking, some of the most current nature-inspired/bio-inspired optimisation approaches produced is the Dragon Fly Algorithm, Moth Flame Optimisation, and Ant Lion Optimisation, Harris Hawk optimisation (HHO), Flower Pollination Algorithm. These algorithms are commonly used to solve complex optimisation problems in a variety of security applications. The DRF is one of the newest optimisation algorithms and has several advantages over previous techniques. It has a low processing cost, less local optimum, rapid convergence, and guards against algorithm stacking during optimisation. Furthermore, the DRF35 is not specifically utilised in applications for IoT-IDS security. Therefore, the goal of the proposed study is to use this method to dataset feature optimisation based on the best optimum solution. Additionally, this optimisation procedure facilitates a simpler classification method with a higher assault detection rate [23].

The balanced IoT dataset's characteristics may be optimally tuned using this optimization approach. Foxes belong to many Canidae families and are tiny to medium-sized omnivore animals with pointed noses, long, thin legs, thick tails, and slender limbs. The foxes may also be distinguished from each other of their family and from large dogs. A novel meta-heuristic optimization system called the DRF takes its cues from the hunting habits of red foxes. When hunting, the red fox moves slowly towards its prey as it hides in the underbrush, and then it attacks the animal out of the blue. Like previous meta-heuristic models, this approach takes into account both the utilization and investigation of capabilities. This method creates random people for initializing parameters, as seen by the subsequent Eq. (3) and Eq. (4).

$$R = [r_0, r_1, \dots, r_{n-1}] \quad (3)$$

$$(R)^i = [(r_0)^i, (r_1)^i, \dots, (r_{n-1})^i] \quad (4)$$

where, "I" denotes how many populations are present in the search area. Ten, the global optimal function is used to find the best solution in the search space. Here, the structure that follows is used in conjunction with the Euclidean distance to get the best solution as presented in Eq. (5).

$$E(((R)^i)^k, (R_{best})^k) = \sqrt{(R^i)^k - (R_{best})^k} \quad (5)$$

In Eq. (5)  $k$  denotes the number of iterations. The term " $R_{best}^t$ " represents the best optimum, while " $E(.)$ " denotes the Euclidean distance. Accordingly, the optimal solution is employed to migrate all candidates, as illustrated in Eq. (6):

$$((R)^i)^k = ((R)^i)^{k-1} + g_{sigm}((R_{best})^k - (R^i)^k) \quad (6)$$

As a scaling hyperparameter, " $g$ " denotes a random value selected at random from 0 to 1 for each iteration. For the whole population, this value is set just once every iteration. People evaluate the fitness values at their new places after moving to the optimal posture. People stay in their new roles if the fitness values are greater; if not, they return to their previous ones. This procedure is similar to how close relatives tell others where to hunt after an adventure and return home. They do what the explorers have instructed, going home "empty-handed" if they don't locate food, or continuing to search if there is a possibility. These processes, which take place during every DRF cycle, resemble suggested global inquiries. In addition, the applicants' move to new roles must present a feasible alternative; if not, their previous jobs will remain. The comparison of the red fox, advancing towards its prey and watches it, is appropriate here since it is similar to the DRF model in which a random number  $\omega$  between 0 and 1 is assumed explained in Eq. (7) and Eq. (8) [24].

$$\begin{cases} \text{Move Forward if, } \omega > \frac{3}{4} \\ \text{Stay Hidden if, } \omega > 3/4 \end{cases} \quad (7)$$

$$\omega = \begin{cases} h \times \frac{\sin(\delta_0)}{\delta_0} & \text{if } \delta_0 \neq 0 \\ \tau & \text{if } \delta_0 = 0 \end{cases} \quad (8)$$

Here, " $h$ " is a random number in the interval  $[0, 0.2]$ , and " $\delta_0$ " is another random number in the interval  $[0, 2\pi]$ , which indicates the fox viewing angle. Furthermore, " $\tau$ " represents a random number between 0 and 1. To model motions for the population of persons, the set of solutions for geographic coordinates is as follows. All things considered, the incorporation of RFO for picking features in intrusion detection systems improves computing efficiency and scalability while also strengthening the system's capacity to precisely detect and address security threats in Internet of Things networks. This method emphasises how crucial it is to use cutting-edge optimisation strategies in order to optimise feature subsets and improve intrusion detection technologies' overall effectiveness.

#### H. Intrusion Detection using Attention Bi-LSTM

The Attention-based BiLSTM model is used to identify intrusions in the NSL-KDD dataset. Using specialised memory units, LSTM—an improved version of the classic Recurrent Neural Networks (RNN)—captures long-term relationships in the MTS dataset efficiently [20]. The gradient vanishing problem is addressed by LSTM models, in contrast to conventional RNN techniques. Rather than depending just on the architecture of hidden units, they also incorporate memory cells that capture the long-term dependence of the signal. Four regulated gates make up the LSTM model: an output gate, a forget gate, input gate, in addition to a self-loop memory cell. These gates control how several memory neurons' data streams communicate with one another. The

forget gate in the LSTM model's hidden layer decides which data from the previous time frame to keep and which to discard. The input gate makes the decision to simultaneously inject data from the memory unit into the input signal or not. The output gate decides whether to change the state of the memory unit [24]. The following Eq. (9) through Eq. (14) are used to determine the neuron state, hidden layer results, and gate states, taking into account the input  $x_t$  from the NSL-KDD dataset and the dynamic output state  $h_t$ :

$$ip_t = \sigma(X_i u_t + Y_i h_{t-1} + a_i) \quad (9)$$

$$fg_t = \sigma(X_f u_t + Y_f h_{t-1} + a_f) \quad (10)$$

$$op_t = \sigma(X_o u_t + Y_o h_{t-1} + a_o) \quad (11)$$

$$c_t = fg_t \odot c_{t-1} + ip_t \odot \tilde{c}_t \quad (12)$$

The weight matrices that recur are indicated by as  $Y_i, Y_f, Y_o$ , while the representation of the weighted matrix for the forget, output, input, and memory cell gating by  $X_i, X_f, X_o$ , respectively. The biases for the gates are formulated as  $a_i, a_f, a_o$ . The candidate's cell state  $\tilde{c}_t$ , is utilized to update the original memory cell state,  $c_t$ . Step indicates the hidden layer's state  $h_{t-1}$  at any given moment, while  $ot$  indicates the output  $op_t$ . The symbol  $\odot$  denotes the element-wise multiplication operation. The hyperbolic tangent function is denoted as  $\tanh$ , and the logistic sigmoid activation function is represented by  $\sigma$ .

The standard LSTM model's limitation lies in its one-directional analysis of input signals during training, potentially leading to the inadvertent oversight of sequential information. In contrast, the BiLSTM was designed with a bidirectional structure, leveraging two LSTM layers operating in opposing directions to capture representation information both forwards and backwards. This bidirectional setup includes a hidden layer for reverse transmission (denoted as  $hb(t)$ ), incorporating future values, alongside a forward propagation hidden layer ( $hf(t)$ ) that retains data from previous sequence values. Ultimately, the BiLSTM model's final output is a fusion of both  $hf(t)$  and  $hb(t)$ , facilitating a more comprehensive understanding of time series data.

$$M_{fg}(t) = \varphi(Y_{fm} u_t + Y_{fmm} u_{f(t-1)} + a_{fa}) \quad (13)$$

$$M_a(t) = \varphi(Y_{am} u_t + Y_{amm} u_{a(t-1)} + a_a) \quad (14)$$

Besides these,  $a_{fa}$  and  $a_a$  also relate to two-way biased data. The weight matrix " $Y_{fm}$  and  $Y_{am}$ " represents the synaptic weights from the input value to the internal unit for both forward and backward directions. Similarly, the forward and backward feedback recurrent weights are denoted by  $Y_{fmm}$  and  $Y_{amm}$ .

The  $\tanh$  function serves as the activation function  $\psi$  for the hidden layers (HLs). It determines the output of the BiLSTM as  $b_t$ .

$$b_t = \sigma(W_{fmb} m_{f(t)} + W_{amb} m_{a(t)} + a_b) \quad (15)$$

The forward and backward weights of the resulting layers are represented by  $W_{fmb}$  and  $W_{amb}$ , respectively, in Eq. (15). Both a linear or sigmoidal function is provided as the



activation function of the resulting layer  $\sigma$ . Moreover,  $b$  denotes the bias in the output. The attention mechanism contributes to the learning process of the Attention BiLSTM model by assigning varying weights. The attention  $a_i$  for a hidden layer  $h_i$  is calculated using Eq. (16):

$$x_i = \tanh(Wh_i + a) \quad (16)$$

BiLSTM networks provide a powerful means to examine sequential data streams, enabling real-time detection of anomalous behavior and security threats in IoT networks. Leveraging BiLSTM architectures, these networks excel in capturing temporal dependencies and patterns present in IoT data, which are often characterized by their dynamic and time-varying nature. By effectively modelling the sequential nature of IoT data, BiLSTM networks can accurately identify deviations from normal behavior, facilitating prompt detection of intrusions and security breaches. To protect the integrity and confidentiality of IoT systems and devices, respond proactively to new threats, and strengthen the security posture of IoT networks, this capability is essential.

### 1. Blockchain Integration

The integration of blockchain technology into intrusion detection systems involves several key steps to ensure the integrity and immutability of the data while facilitating secure communication among distributed IoT devices through smart contracts.

1) *Data logging*: In the process of data logging, intrusion detection data generated by IoT devices is systematically recorded onto the blockchain network. Each piece of data is meticulously timestamped and cryptographically secured, ensuring its integrity and safeguarding against any potential tampering attempts. By timestamping each entry, the blockchain network establishes a chronological order of events, enabling a comprehensive audit trail of intrusion activities. Additionally, the cryptographic security measures implemented within the blockchain network guarantee the immutability of the logged data, thereby providing a reliable and tamper-proof record of security events. This meticulous logging process enhances the trustworthiness and reliability of the intrusion detection system, enabling robust security monitoring in IoT networks [25].

2) *Blockchain node*: In the context of blockchain technology, blockchain nodes serve as essential components responsible for validating and recording logged intrusion detection data. These nodes are distributed across the blockchain network, ensuring decentralization and resilience against single points of failure. Each node maintains a copy of the decentralized ledger, which contains a complete record of all transactions, including the logged intrusion detection data. When new data is logged onto the blockchain, it undergoes validation by multiple nodes within the network to ensure its authenticity and integrity. This validation process involves verifying the cryptographic signatures associated with the data and confirming its adherence to the consensus rules established by the network protocol. Once validated, the intrusion detection data is appended to the blockchain ledger,

becoming a permanent and immutable part of the distributed database. By distributing the responsibility for data validation and storage among multiple nodes, blockchain networks achieve redundancy and fault tolerance, enhancing the reliability and resilience of the overall system. Furthermore, as blockchain nodes are decentralised, no one organisation can exert control over the system as a whole, fostering openness, confidence, and security in the logging and archiving of intrusion detection data.

3) *Proof of work*: The consensus mechanism of the blockchain is essential to guaranteeing that all dispersed nodes agree on the veracity of logged data. To reach this consensus among network users, consensus techniques like Proof of Work (PoW) are used. Proof-of-work (PoW) consensus is a competitive mechanism in which nodes solve challenging mathematical problems to validate transactions and append new blocks to the blockchain. This is a resource-intensive procedure that uses a lot of energy and processing power. Nonetheless, other nodes in the network confirm the answer after a node completes the puzzle and suggests a new block. The block is appended to the blockchain if the answer satisfies the consensus requirements. By using this decentralised method, blockchain networks maintain the integrity and durability of the blockchain ledger by facilitating consensus across dispersed nodes about the veracity of recorded data. Additionally, consensus mechanisms like PoW contribute to the security of the blockchain network by mitigating the risk of malicious actors attempting to manipulate or alter the logged data. Overall, the consensus mechanism serves as a fundamental building block of blockchain technology, enabling decentralized trust and coordination among network participants [26].

A key element of blockchain networks is the proof-of-work (PoW) consensus mechanism, which guarantees dispersed nodes' agreement on the legitimacy of transactions and the appending of new blocks to the blockchain. PoW comprises the following crucial steps:

- **Transaction Propagation**: Transactions are broadcasted to all nodes in the blockchain network. Each transaction contains details such as sender, recipient, amount, and cryptographic signatures.
- **Block Creation**: Transactions are grouped together into blocks, forming a candidate block for addition to the blockchain. Miners, who are nodes responsible for creating new blocks, select transactions and assemble them into a block structure.
- **Mining Competition**: Miners compete with each other to solve the Proof of Work puzzle. They utilize computational power to generate hash values by iteratively modifying a nonce (a random number) in the block header until the desired hash value is found. This process is computationally intensive and requires significant computational resources.

- **Verification:** A miner broadcasts the candidate block and the solution to the network as soon as they discover a workable solution to the problem. The legitimacy of the answer and the transactions included in the block are then confirmed by further nodes inside the network.
- **Consensus:** If the majority of nodes in the network agree that the proposed solution is sound and the block conforms to the consensus requirements, the block is accepted and posted to the blockchain. It is ensured that all distributed nodes concur on the validity of the transactions and the addition of new blocks to the blockchain by going through this process.
- **Reward:** A fixed quantity of bitcoin plus any transaction fees included in the block are awarded to the miner who effectively mines a new block. This encourages miners to use up processing power and take part in the consensus-building process on the network.

In general, the Proof of Work technique reduces the possibility of malevolent actors attempting to influence the blockchain by demanding computational resources to verify transactions and generate new blocks, hence ensuring the security and integrity of blockchain networks.

1) *Smart contract:* Smart contracts serve as the backbone of automation and governance within IoT networks by providing a decentralized, programmable framework for enforcing rules and conditions. These contracts, encoded with predefined logic, are deployed on the blockchain, ensuring immutability and tamper-proof execution. Within the context of IoT, smart contracts automate interactions between devices, enabling seamless communication and coordination without the need for intermediaries. By executing automatically when specific conditions are met, such as sensor readings or trigger events, smart contracts streamline processes and mitigate the risk of human error. Moreover, the decentralised structure of these systems gets rid of single points of failure and minimises dependence on centralised authority, hence improving security and resilience. Additionally, conditional execution of operations is made possible by smart contracts, which let gadgets react quickly to shifting conditions. This feature improves IoT network responsiveness and operational efficiency. Furthermore, network participants' confidence and responsibility are bolstered by the openness and auditability provided by smart contracts. Overall, smart contracts play a critical role in driving efficiency, security, and transparency in IoT ecosystems, laying the foundation for scalable and resilient decentralized applications [27].

2) *Secure communication:* In the ecosystem of IoT networks, secure communication is facilitated through the interaction between IoT devices and the blockchain network via smart contracts. These contracts act as intermediaries, enforcing cryptographic protocols and access controls to ensure that communication remains secure. By leveraging cryptographic techniques such as encryption and digital signatures, smart contracts authenticate and authorize devices,

mitigating the risk of unauthorized access or tampering. Through predefined rules and conditions encoded within the smart contracts, only authorized devices are granted permission to access and modify data stored on the blockchain. This robust enforcement of security measures enhances the integrity and confidentiality of communication within IoT networks, safeguarding sensitive information and preventing unauthorized manipulation of data. Overall, the utilization of smart contracts enables secure and trustworthy communication channels, fostering confidence in the exchange of data and transactions within IoT ecosystems.

## V. RESULT AND DISCUSSION

The proposed framework undergoes rigorous evaluation using benchmark datasets, including NSL-KDD and BoT-IoT, to comprehensively assess its performance in detecting various types of intrusions within IoT networks. By leveraging these datasets, which contain diverse and realistic intrusion scenarios, the framework's efficacy in identifying and mitigating security threats is thoroughly scrutinized. Performance metrics are used to assess how well the framework differentiates between malicious activity and typical network behavior. These measures include detection accuracy, false positive rate, and computing efficiency. Furthermore, the assessment procedure entails contrasting the outcomes of the framework with those of current intrusion detection systems in order to measure its effectiveness in relation to predetermined benchmarks. The suggested framework's potential to strengthen the security posture of IoT networks is carefully investigated through this methodical study utilizing typical datasets, offering insights into its advantages and shortcomings.

### A. Performance Metrics

Performance metrics refer to the numerical values that are utilized to assess how well an intrusion detection system detects and neutralizes security threats on a network. Commonly used metrics include the following ones:

1) *Accuracy:* The percentage of accurately identified occurrences—both true positives and true negatives—out of all the instances that were examined is known as accuracy. It offers a general indicator of how effectively the intrusion detection system classifies events as either intrusions or routine activity.

2) *Precision:* Positive predictive value, or precision, is a metric that expresses the percentage of accurately detected positive cases (true positives) across every case categorized as positive (false positives and true positives). It shows how well the system can detect intrusions without mistakenly labelling routine operations as such.

3) *Recall:* Recall, also known as sensitivity or true positive rate, is the proportion of correctly identified positive cases relative to all real positive occurrences in the dataset. It assesses the system's ability to identify every incursion, lowering the likelihood that any malicious activity would go undetected.

4) *F1-score*: The F1-score, which achieves equilibrium between recall and accuracy, is derived from the harmonic mean of these two metrics. Recall and accuracy are combined into one figure, which accounts for both false positives and false negatives.

TABLE I. PERFORMANCE METRICS

Metrics	Efficiency
Accuracy	98.9
Precision	94
Recall	95
F1-Score	95

As shown in Table I and Fig. 2, the suggested intrusion detection approach exhibits excellent efficiency with an accuracy of 98.9%, demonstrating its capacity to accurately categorise cases as either intrusions or routine operations. Furthermore, the approach displays a 94% accuracy rate, which indicates the percentage of accurately detected incursions among all cases that are categorised as positive, hence reducing false positives. With a recall rate of 95%, which indicates that the system can detect all incursions, there is little chance of a missed detection. Furthermore, a balanced performance in terms of both accuracy and recall is shown by the F1-score, which harmonises the two metrics, which is recorded at 95%. All of these measures show how successful and dependable the suggested intrusion detection technique is at identifying and reducing security risks in the network infrastructure.



Fig. 2. Performance efficiency.

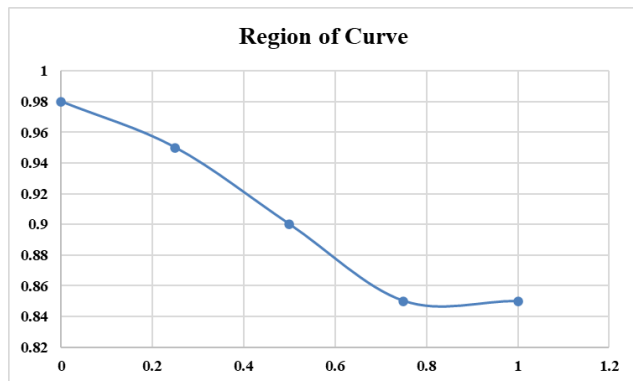


Fig. 3. Receiver operating characteristic curve.

As the threshold rises from 0 to 1, the true positive rate (TPR) progressively falls from 0.98 to 0.85, suggesting a decline in the percentage of true positive cases that are correctly categorised, as seen in Fig. 3. The TPR stays comparatively high at 0.95 at a threshold of 0.25, indicating that true positive cases can be effectively detected even with somewhat loosened thresholds.

TABLE II. SORTING RESULT OF NSL-KDD

Methods	AUC	Error Rate
Gradient Boosting Classifier	47.64	0.4905
Deep Learning	77.88	0.2256
Proposed Method	98.9	0.0025

The NSL-KDD dataset's categorization outcomes using different techniques are shown in Table II. With an error rate of 0.4905 and an AUC of 47.64%, the Gradient Boosting Classifier performs relatively poorly. By comparison, the Deep Learning approach shows noticeably higher performance, with an error rate of 0.2256 and an AUC of 77.88%.

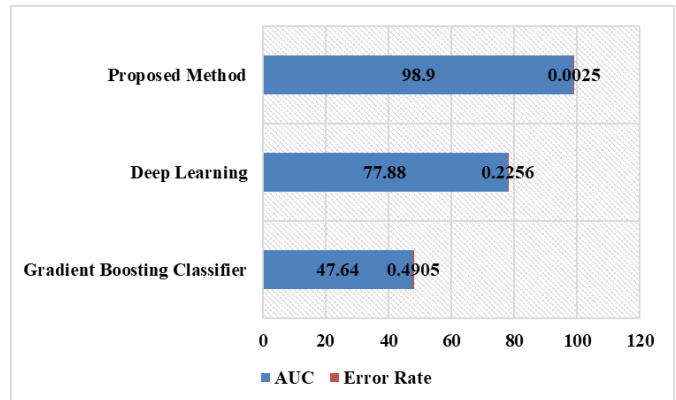


Fig. 4. Classification result of NSL-KDD.

The results presented in Fig. 4 demonstrate that the suggested approach outperforms the two other options, with an exceptional AUC of 98.9% and a remarkably low error rate of 0.0025. These outcomes highlight how well the suggested strategy performs in comparison to other methods when it comes to correctly identifying instances in the NSL-KDD dataset.

TABLE III. RECOGNITION OUTCOMES OF ATTENTION BASED BiLSTM APPROACH ON NSL-KDD DATASET

Data	Class	Accuracy	Precision	Recall	F1-Score
Training	Normal	98.4	96.3	97.3	96.3
	Attack	97.4	97.4	98.4	95.5
	Average	97.7	97.7	97.7	97.7
Testing	Normal	98.9	97.5	96.4	95.8
	Attack	97.3	98.3	97.3	98.3
	Average	98.9	98.9	98.9	98.9

The NSL-KDD dataset's recognition results from the Attention-based BiLSTM technique are shown in Table III and Fig. 5.

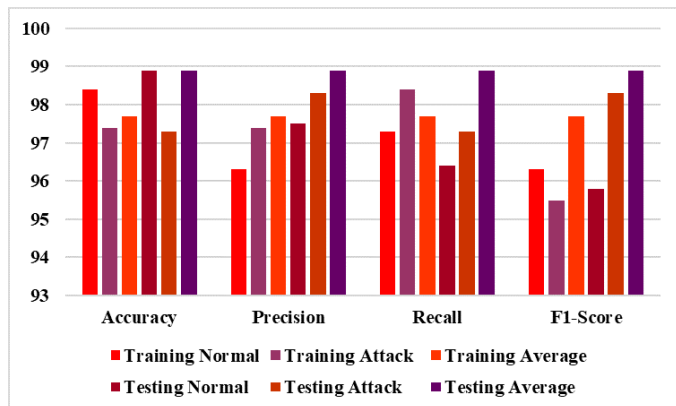


Fig. 5. Recognition outcomes of attention based BiLSTM approach on NSL-KDD dataset.

It includes accuracy, precision, recall, and F1-Score, split into normal and attack classes, for both the training and testing datasets. In the training dataset, the method achieves 98.4% accuracy for normal cases and 97.4% accuracy for attack instances. Respectively, the corresponding accuracy, recall, and F1-Score values are 95.5% and 96.3%, 97.3% and 98.4%, and 96.3% and 97.4%. Comparable outcomes are seen in the testing dataset, where the technique achieves 97.3% accuracy for attack instances and 98.9% accuracy for normal cases. The corresponding F1-Score, recall, and accuracy scores are 97.5%, 98.3%, and 95.8%, respectively. The average results for each class are also provided for the training and testing datasets.

### B. Discussion

The research studies under discussion offer novel strategies for resolving the security issues that arise in Internet of Things networks. These specifically concentrate on intrusion detection systems (IDS) and make utilisation of blockchain and machine learning technology. The study by Liang et al. [12] suggests a hybrid intrusion detection system that makes use of multi-agent systems, blockchain technology, and deep learning techniques. The system is divided into distinct modules for data collecting, management, analysis, and response with the goal of improving detection accuracy, particularly at the transport layer of Internet of Things networks. Scalability and optimisation continue to be major obstacles to practical implementation, notwithstanding encouraging findings. A collaborative intrusion detection architecture including blockchain technology for safe sharing of threat intelligence across cloud and Internet of Things networks is presented by Alkadi et al. [13]. While consensus processes and deep blockchain technology are adept at detecting intrusions and reducing security threats, their scale presents serious problems for efficiency and real-time response. A blockchain-based collaborative signature-based IDS called CBSigIDS is proposed by Li et al. with the goal of creating a trustworthy signature database in dispersed IoT systems. Although it provides a safe way to validate signatures, blockchain overhead scalability issues require

further work before a viable implementation can be made. Kumar et al. [14] offers a distributed intrusion detection system (IDS) that uses blockchain technology and fog computing to identify DDoS assaults directed at IoT mining pools. They assess the system's effectiveness in identifying IoT network assaults using machine learning algorithms trained on scattered fog nodes. But there are still issues with realistic implementation and optimisation needed for efficiency and scalability. Although these studies show how blockchain and machine learning technologies could potentially use to improve IoT network security, scalability, optimisation, and practical deployment issues must be resolved before their full promise could be realised in practical settings.

The study presents a complete framework for reliable and scalable intrusion detection in IoT networks by integrating machine learning techniques with blockchain technology. The solution addresses the challenges posed by the dynamic and heterogeneous nature of IoT environments by employing Red Fox Optimization for feature selection and Attention-based BiLSTM for anomaly identification. The adoption of blockchain technology improves security by ensuring the validity and inviolability of intrusion record detection. The study advances the area by providing an all-encompassing method of intrusion detection that takes security and efficiency into account. Real-time identification of abnormalities and malicious activity in IoT traffic is made possible by the use of sophisticated machine learning algorithms, and scalability is improved by optimization approaches that assist decrease the dimensionality of the input data. Furthermore, the system gains an additional degree of protection through the integration of the technology known as blockchain, which offers tamper-resistant recordings of detected intrusions. The usefulness of the suggested architecture is demonstrated by experimental findings, which on real-world IoT datasets yield a high detection accuracy of about 98.9%. These findings highlight how important the study is to improving IoT security state-of-the-art. The report does, however, admit several limitations, including the need for more assessment in various IoT scenarios and the computational cost related to blockchain integration. Prospective study avenues encompass investigating alternative machine learning algorithms and optimization methods, tackling scalability issues, and refining blockchain-associated procedures. Overall, the research offers a viable strategy for improving intrusion detection in Internet of Things networks, opening the door to more robust and safe linked settings.

## VI. CONCLUSION

The suggested system, which makes use of blockchain and machine learning, offers a viable solution to the problems associated with intrusion detection in Internet of Things networks. The accuracy and scalability of the intrusion detection system are improved by integrating Red Fox Optimization for feature selection and Attention-based BiLSTM for anomaly detection. Moreover, the incorporation of blockchain technology ensures the integrity and immutability of intrusion detection logs, thereby enhancing security. On real-world IoT data sets, experimental findings show the usefulness of the technique with a high detection

accuracy of about 98.9%. However, it is important to acknowledge some limitations and areas for future work. Firstly, while the proposed framework shows promising results, further research is needed to evaluate its performance in diverse IoT environments and under various attack scenarios. Additionally, the scalability of the system needs to be investigated to handle large-scale IoT networks efficiently. Furthermore, the computational overhead associated with blockchain integration may pose challenges in resource-constrained IoT devices, requiring optimization strategies. Moreover, continuous advancements in intrusion techniques necessitate ongoing updates and improvements to the detection algorithms and feature selection methods. Future studies may look at applying more machine learning algorithms and optimization techniques to enhance the robustness and efficiency of intrusion detection systems in Internet of Things networks. All things considered, this work establishes the groundwork for next investigations that seek to create IoT ecosystems that are more robust and safer.

#### REFERENCES

- [1] P. Raj and A. C. Raman, *The Internet of Things: Enabling technologies, platforms, and use cases*. Auerbach Publications, 2017.
- [2] V. E. Balas and S. Pal, *Healthcare Paradigms in the Internet of Things Ecosystem*. Academic Press, 2020.
- [3] Y. Liao, C. Thompson, S. Peterson, J. Mandrola, and M. S. Beg, "The future of wearable technologies and remote monitoring in health care," *Am. Soc. Clin. Oncol. Educ. Book*, vol. 39, pp. 115–121, 2019.
- [4] A. Karale, "The challenges of IoT addressing security, ethics, privacy, and laws," *Internet Things*, vol. 15, p. 100420, 2021.
- [5] A. Qasem, P. Shirani, M. Debbabi, L. Wang, B. Lebel, and B. L. Agba, "Automatic vulnerability detection in embedded devices and firmware: Survey and layered taxonomies," *ACM Comput. Surv. CSUR*, vol. 54, no. 2, pp. 1–42, 2021.
- [6] R. Krishnamurthi, A. Kumar, D. Gopinathan, A. Nayyar, and B. Qureshi, "An overview of IoT sensor data processing, fusion, and analysis techniques," *Sensors*, vol. 20, no. 21, p. 6076, 2020.
- [7] A. Riah, S. Daniel, E. Frank, and K. Seriffdeen, "The role of technology in shaping user behavior and preventing phishing attacks," 2024.
- [8] T. M. Alshammari and F. M. Alserhani, "Scalable and Robust Intrusion Detection System to Secure the IoT Environments using Software Defined Networks (SDN) Enabled Architecture," *Int J Comput Netw. Appl.*, vol. 9, no. 6, pp. 678–688, 2022.
- [9] M. Javed, N. Tariq, M. Ashraf, F. A. Khan, M. Asim, and M. Imran, "Securing Smart Healthcare Cyber-Physical Systems against Blackhole and Greyhole Attacks Using a Blockchain-Enabled Gini Index Framework," *Sensors*, vol. 23, no. 23, p. 9372, 2023.
- [10] A. Laszka, A. Dubey, M. Walker, and D. Schmidt, "Providing privacy, safety, and security in IoT-based transactive energy systems using distributed ledgers," in *Proceedings of the Seventh International Conference on the Internet of Things*, 2017, pp. 1–8.
- [11] A. K. Al Hwaitat et al., "A New Blockchain-Based Authentication Framework for Secure IoT Networks," *Electronics*, vol. 12, no. 17, p. 3618, Aug. 2023, doi: 10.3390/electronics12173618.
- [12] C. Liang et al., "Intrusion Detection System for the Internet of Things Based on Blockchain and Multi-Agent Systems," *Electronics*, vol. 9, no. 7, p. 1120, Jul. 2020, doi: 10.3390/electronics9071120.
- [13] O. Alkadi, N. Moustafa, B. Turnbull, and K.-K. R. Choo, "A deep blockchain framework-enabled collaborative intrusion detection for protecting IoT and cloud networks," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9463–9472, 2020.
- [14] R. Kumar, P. Kumar, R. Tripathi, G. P. Gupta, S. Garg, and M. M. Hassan, "A distributed intrusion detection system to detect DDoS attacks in blockchain-enabled IoT network," *J. Parallel Distrib. Comput.*, vol. 164, pp. 55–68, Jun. 2022, doi: 10.1016/j.jpdc.2022.01.030.
- [15] H. Vargas, C. Lozano-Garzon, G. A. Montoya, and Y. Donoso, "Detection of Security Attacks in Industrial IoT Networks: A Blockchain and Machine Learning Approach," *Electronics*, vol. 10, no. 21, p. 2662, Oct. 2021, doi: 10.3390/electronics10212662.
- [16] R. H. Hylock and X. Zeng, "A Blockchain Framework for Patient-Centered Health Records and Exchange (HealthChain): Evaluation and Proof-of-Concept Study," *J. Med. Internet Res.*, vol. 21, no. 8, p. e13592, Aug. 2019, doi: 10.2196/13592.
- [17] "NSL-KDD." Accessed: Mar. 21, 2024. [Online]. Available: <https://www.kaggle.com/datasets/hassan06/nslkdd>.
- [18] H. Moudoud, S. Cherkaoui, and L. Khoukhi, "An IoT blockchain architecture using oracles and smart contracts: the use-case of a food supply chain," in *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*, IEEE, 2019, pp. 1–6.
- [19] P. Bari and P. Karande, "Application of PROMETHEE-GAIA method to priority sequencing rules in a dynamic job shop for single machine," *Mater. Today Proc.*, vol. 46, pp. 7258–7264, 2021, doi: 10.1016/j.matpr.2020.12.854.
- [20] P. Yazdaniyan and S. Sharifian, "E2LG: a multiscale ensemble of LSTM/GAN deep learning architecture for multistep-ahead cloud workload prediction," *J. Supercomput.*, vol. 77, pp. 11052–11082, 2021.
- [21] F. Karim, S. Majumdar, and H. Darabi, "Insights Into LSTM Fully Convolutional Networks for Time Series Classification," *IEEE Access*, vol. 7, pp. 67718–67725, 2019, doi: 10.1109/ACCESS.2019.2916828.
- [22] Z. Ahamed, M. Khemakhem, F. Eassa, F. Alsolami, and A. S. A.-M. Al-Ghamdi, "Technical Study of Deep Learning in Cloud Computing for Accurate Workload Prediction," *Electronics*, vol. 12, no. 3, p. 650, 2023.
- [23] E. S. P. Krishna and A. Thangavelu, "Attack detection in IoT devices using hybrid metaheuristic lion optimization algorithm and firefly optimization algorithm," *Int. J. Syst. Assur. Eng. Manag.*, May 2021, doi: 10.1007/s13198-021-01150-7.
- [24] R. AlGhamdi, "Design of Network Intrusion Detection System Using Lion Optimization-Based Feature Selection with Deep Learning Model," *Mathematics*, vol. 11, no. 22, p. 4607, Nov. 2023, doi: 10.3390/math11224607.
- [25] S. Rathore, J. H. Park, and H. Chang, "Deep Learning and Blockchain-Empowered Security Framework for Intelligent 5G-Enabled IoT," *IEEE Access*, vol. 9, pp. 90075–90083, 2021, doi: 10.1109/ACCESS.2021.3077069.
- [26] A. H. Sodhro, S. Pirbhulal, M. Muzammal, and L. Zongwei, "Towards blockchain-enabled security technique for industrial internet of things based decentralized applications," *J. Grid Comput.*, vol. 18, pp. 615–628, 2020.
- [27] B. Yin, Y. Wu, T. Hu, J. Dong, and Z. Jiang, "An efficient collaboration and incentive mechanism for Internet of Vehicles (IoV) with secured information exchange based on blockchains," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1582–1593, 2019.

Original Article

# Advancing Real-Time Pedestrian Behavior Analysis at Zebra Crossings with Transfer Learning and Pre-trained Model

Pannalal Boda<sup>1</sup>, Y. Ramadevi<sup>2</sup>

<sup>1</sup>Department of CSE, Osmania University, Hyderabad, Telangana, India.

<sup>2</sup>Department of CSE, Chaitanya Bharathi Institute of Technology, Osmania University, Hyderabad, Telangana, India.

<sup>2</sup>Corresponding Author : [yramadevi\\_cse@cbit.ac.in](mailto:yramadevi_cse@cbit.ac.in)

Received: 25 March 2024

Revised: 16 May 2024

Accepted: 25 May 2024

Published: 29 June 2024

**Abstract** - This paper introduces the Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework, aimed at enhancing real-time pedestrian behavior analysis at zebra crossings to improve urban traffic safety and facilitate the integration of autonomous vehicles. Addressing limitations in real-time applicability, accuracy under diverse conditions, and scalability of current methodologies, the PPASE utilizes transfer learning and pre-trained models tailored for pedestrian behavior. Leveraging the Pedestrian Intention Estimation (PIE) dataset, enriched with real-time urban traffic data, the framework offers refined predictions of pedestrian movements. Performance is rigorously evaluated using accuracy, precision, recall, and F1 score, with the PPASE demonstrating commendable overall accuracy of 92.5% in pedestrian crossing predictions, 89.4% in movement pattern identification, and 93.7% in group dynamics analysis. These quantitative results highlight the framework's potential to significantly mitigate incidents at zebra crossings and improve crowd management in urban settings, affirming its efficacy as an advanced tool for enhancing pedestrian safety within intelligent urban traffic systems.

**Keywords** - Pedestrian Behavior Analysis, Urban Traffic Safety, Autonomous vehicles, Real-Time Prediction, Group dynamics.

## 1. Introduction

The study of pedestrian behavior in urban settings has evolved significantly over the past few decades, spurred by the increasing need to improve road safety, manage traffic flow, and integrate autonomous vehicles into urban environments. Initially, pedestrian behavior analysis relied heavily on observational studies and manual data collection methods. These early approaches provided valuable insights into pedestrian dynamics but were time-consuming, labor-intensive, and limited in scope and scalability. With the advent of digital technology and computing power, the 1990s and early 2000s saw a shift towards automated surveillance systems and the use of computer vision techniques [1].

These advancements allowed for more comprehensive data collection and analysis, enabling researchers to study pedestrian behavior in greater detail and over larger areas. Computer vision techniques, such as object detection and tracking, became foundational in understanding pedestrian movements and interactions in urban spaces. The proliferation of machine learning and artificial intelligence in the last decade has further transformed pedestrian behavior analysis [2]. Researchers have begun to apply sophisticated machine learning models, including deep learning, to predict pedestrian actions with greater accuracy. These models can process vast

amounts of data from diverse sources, such as CCTV footage, smartphone sensors, and GPS data, to learn complex patterns of pedestrian behavior [3]. Moreover, simulation technologies have also played a crucial role, enabling researchers to model and predict pedestrian movements under various scenarios and conditions.

These simulations help in understanding the impact of different urban designs and traffic management strategies on pedestrian safety and traffic efficiency. Despite these technological advancements in the evolving landscape of urban mobility, the safety and efficiency of road traffic are paramount, necessitating accurate predictions of pedestrian behavior at zebra crossings [4]. This is crucial for urban planning, traffic management, and the integration of autonomous vehicle systems. Accurate behavior prediction aids in designing safer urban environments, optimizing traffic flow, and ensuring the safety of all road users. However, the challenge of real-time pedestrian behavior analysis is magnified by environmental variability and the diversity of pedestrian actions, necessitating swift, precise predictions. Existing methodologies, including computer vision, machine learning, and simulation techniques, provide foundational insights but struggle with real-time applicability, accuracy under diverse conditions, and scalability.



Recent studies underscore the limitations of traditional methods, highlighting the occurrence of serious injuries or fatalities at zebra crossings due to risky behaviors, such as mobile phone usage while crossing. These findings point to a critical need for improved analytical techniques capable of real-time operation to enhance urban traffic safety and efficiency [5]. Addressing these challenges, our study leverages transfer learning and fine-tuning of pre-trained models, promising approaches in the domain of image recognition and behavior prediction. By adapting these models with pedestrian-specific data, we aim to develop a scalable, robust solution for real-time pedestrian behavior analysis, enhancing the overall safety of urban traffic systems [6]. The motivation behind this research is the imperative need to improve pedestrian safety and traffic management through advanced predictive analytics. The key contributions of this paper include the development of an efficient framework for real-time pedestrian behavior prediction at zebra crossings, significantly improving accuracy in diverse conditions, and offering a robust solution for urban traffic systems and autonomous vehicle navigation.

Key contributions of the research paper are

1. Developed the Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework, leveraging transfer learning and pre-trained models for the real-time prediction of pedestrian behaviors at zebra crossings, thereby improving urban traffic safety and management.
2. Developed a sophisticated solution for analyzing pedestrian behaviors, encompassing crossing actions, movement patterns, and group dynamics, significantly enhancing traffic safety measures in diverse urban environments.
3. The research paper significantly contributes by rigorously validating the proposed model's effectiveness using critical evaluation metrics, including accuracy, precision, recall, and F1 score.

The remainder of this paper is organized as follows: Section 2 delves into the literature review, offering a comprehensive overview of relevant studies and existing knowledge. Section 3 introduces the proposed methodology, detailing the approach and techniques employed in this research. Section 4 discusses the results and analysis, providing insights into the findings and their implications. Finally, Section 5 concludes the paper, summarizing key points and suggesting directions for future research.

## 2. Literature Review

The intersection of pedestrian behavior analysis and urban traffic safety management has garnered considerable attention in recent years, driven by the imperative to mitigate pedestrian-vehicle incidents in urban settings. This literature review synthesizes seminal and contemporary studies within this domain, setting the stage for the contributions of the present research.

Early studies in pedestrian behavior analysis primarily focused on observational techniques to understand pedestrian movements and crossings. These studies faced challenges in terms of time commitment and resource-intensive spatial analysis. However, recent advancements in technology, such as Unmanned Aerial Vehicles (UAVs) and smart transportation systems, have provided new opportunities to study pedestrian behavior. UAV-based observation techniques have shown promise in measuring pedestrian activity, allowing for larger surface area coverage in less time [7]. Additionally, smart transportation systems offer innovative techniques to connect pedestrians, vehicles, and infrastructure, enhancing mobility and safety [8].

Furthermore, studies have utilized video recordings and trajectory data to analyze pedestrian crossing behavior, employing methods like the Kalman filter and topic modeling to understand pedestrian intentions and strategies [9]. These advancements have expanded the scope of pedestrian behavior analysis beyond traditional observational techniques, enabling a more comprehensive understanding of pedestrian movements and crossings. In this article [10], the pedestrian crossing was influenced by a number of factors, which were the most important of which are the time and speed of pedestrian crossings, which are direct dependence on the width of the marked pedestrian crossing. The authors [11] established a classification system for pedestrian interactive behaviors and utilized pose estimation to acquire 2D key points on the skeleton of pedestrians. This approach is used to represent high-level spatio-temporal characteristics based on body pose.

With advancements in technology, recent works have focused on using computational models to improve the accuracy of predicting pedestrian behavior. These models utilize deep learning approaches, such as Convolutional Neural Networks (CNN) and Transformer architectures, to capture the complex interactions and contextual elements that influence pedestrian behavior. For example, a novel framework proposed by Zhang et al. combines a cross-modal Transformer architecture with semantic attentive interaction modules to predict future trajectories and crossing actions of pedestrians [12]. Another study by Deokar and Khandekar explores the use of CNNs to recognize the direction of pedestrian movement, achieving high accuracy in binary and multiclass classification tasks [13]. These advancements in computational models have shown promising results in improving the accuracy and reliability of pedestrian behavior prediction, which is crucial for applications such as autonomous driving systems and pedestrian analysis [14].

Real-time predictive capability is often lacking in existing models for immediate application in traffic safety management [15]. However, recent research has focused on developing models that incorporate real-time data and deep learning techniques to improve prediction accuracy and enable

immediate application [16]. For example, a web-based proactive traffic safety management system has been developed, which utilizes real-time data such as traffic, weather, and video data to predict crashes in real-time. Another study proposes a two-stage framework that combines machine learning algorithms and real-time traffic and weather variables to predict traffic levels and recovery time after an accident. Additionally, transfer-learning approaches have been used to improve the spatiotemporal transferability of deep-learning crash likelihood prediction models, allowing for accurate predictions in new locations. These advancements in real-time predictive models contribute to the improvement of traffic safety management systems.

Transfer learning and pre-trained models have significantly advanced the field of pedestrian behavior prediction [17]. Recent research has shown that pre-training on unlabeled person images leads to superior performance in person re-identification tasks compared to pre-training on ImageNet [18]. However, these pre-trained methods are often designed specifically for re-identification and struggle to adapt to other pedestrian analysis tasks. To address this, novel frameworks like VAL-PAT have been proposed, which learn transferable representations to enhance various pedestrian analysis tasks using multimodal information. Additionally, the use of multitask sequence to sequence Transformer encoders-decoders architectures has been introduced for pedestrian action and trajectory prediction, achieving improved accuracy compared to existing LSTM-based models. These advancements in transfer learning and pre-trained models have greatly contributed to the evolution of pedestrian behavior prediction.

Understanding complex pedestrian behaviors, such as movement patterns and group dynamics, poses a challenge for traditional analytical frameworks. Deep learning-based approaches have gained popularity in recent years due to their superior performance in predicting pedestrian behavior in complex scenarios compared to traditional approaches such as social force or constant velocity models [19]. Additionally, a behavioral model based on Voronoi and Delaunay diagrams has been proposed to deconstruct pedestrian crowds and reproduce realistic motion in simulations, capturing the natural correlation between movement choices and human behaviors [12]. Furthermore, a method combining preprocessing, feature extraction, and CNN classification has been developed to identify anomalous and normal pedestrian behavior, achieving higher performance compared to other approaches [20]. These advancements in deep learning, behavioral modeling, and feature extraction techniques contribute to a better understanding of pedestrian behaviors and can be applied to crowd management and robot navigation.

The contributions of this research paper address the identified gaps by developing the Predictive Pedestrian

Analytics for Safety Enhancement (PPASE) framework. PPASE leverages transfer learning and pre-trained models [21] for the real-time prediction of pedestrian behaviors, offering a sophisticated solution to analyze complex pedestrian dynamics.

Furthermore, this study rigorously validates the PPASE framework's effectiveness using comprehensive evaluation metrics, thereby advancing the state-of-the-art in pedestrian safety enhancement. By situating the PPASE framework within the extant scholarly discourse, this research underscores the novelty and significance of its contributions to the field of urban traffic safety management. The following sections detail the methodology, implementation, and validation of the PPASE framework, highlighting its potential to transform pedestrian safety strategies in urban environments.

### 3. Methodology: Predictive Pedestrian Analytics for Safety Enhancement (PPASE)

In our pursuit to bolster urban traffic safety, the development of an analytical framework that can effectively identify and categorize pedestrian behaviors at zebra crossings stands as a critical endeavor. The Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework is a pioneering initiative in this direction. It capitalizes on the comprehensive and diverse dataset provided by the Pedestrian Intention Estimation (PIE), which encapsulates a wide spectrum of pedestrian behaviors observed in various urban environments. The richness and the annotated nature of the integration of Pedestrian Intention Estimation datasets into the PPASE framework are pivotal for enhancing the prediction of pedestrian behaviors and intentions at zebra crossings. These datasets provide a rich source of pre-analyzed pedestrian behaviors, which are crucial for training the framework's machine-learning models with a focus on intention prediction. PIE Dataset furnish an invaluable asset for conducting detailed analyses of pedestrian actions and their interactions at zebra crossings. Such analyses are instrumental in unraveling the complexities of pedestrian dynamics, serving as a bedrock for predictive modeling and behavioral insights. The PPASE framework is distinguished by its adoption of cutting-edge analytics, leveraging the potent capabilities of transfer learning and pre-trained models. This innovative approach facilitates the real-time prediction of pedestrian behavior with an unprecedented level of accuracy.

By integrating these advanced analytical methodologies, PPASE aims to significantly improve urban traffic safety and management. Its core mission is not just to predict pedestrian movements but to understand the underlying patterns and decision-making processes that govern these movements at zebra crossings. Through this understanding, PPASE endeavors to introduce a paradigm shift in how urban traffic systems accommodate and interact with pedestrians, ensuring a safer and more harmonious coexistence.



**3.1. PPASE Framework: Comprehensive Workflow and Component Functionalities**

The PPASE framework is architected to enhance pedestrian safety at zebra crossings through the collection of real-time data, enriched by integrating Pedestrian Intention Estimation datasets. Leveraging advanced machine learning technologies, including transfer learning and pre-trained models, this framework integrates various components, each dedicated to specific functionalities from data collection to decision support and alert generation. Figure 1 describes an advanced system called the Intention-Aware Pedestrian Analytic System (IAPAS), which is essentially a smart setup for understanding what pedestrians are likely to do next at crosswalks. Imagine a busy city street corner with a crosswalk where our system watches over pedestrians using cameras and sensors. The system starts by collecting all this visual and sensor data, which might include things like where the pedestrians are, how fast they're moving, and in which direction. This collected data is known as the PIE dataset.

The system then goes through a series of steps to make sense of this data. First, it merges the new information with any existing data it has, like previous crosswalk recordings, to

get a fuller picture. Then, it takes a closer look at the details, enhancing key features like how a person is standing or moving to better understand their behavior. After that, it labels these observations with intentions, such as “about to cross” or “just waiting,” which is crucial for the system to learn from past behavior.

Next, the system uses a method called Transfer Learning, which is like giving it a head start with what it already knows from similar tasks and adapting this knowledge specifically for understanding pedestrian movements. It uses something called T-GCN, which helps the system keep track of how pedestrians move over time, not just in a single moment.

Plus, it has a special focus feature that pays extra attention to the most important movements or behaviors that indicate what a person might do next. All this analyzed data is then processed by a part called the Dynamic Intention Insight Framework (DIIF), which does three main things: it looks for patterns in how pedestrians behave, it adds in extra information like the time of day or weather conditions, and finally, it combines all this to make a good guess about what each pedestrian is likely to do.

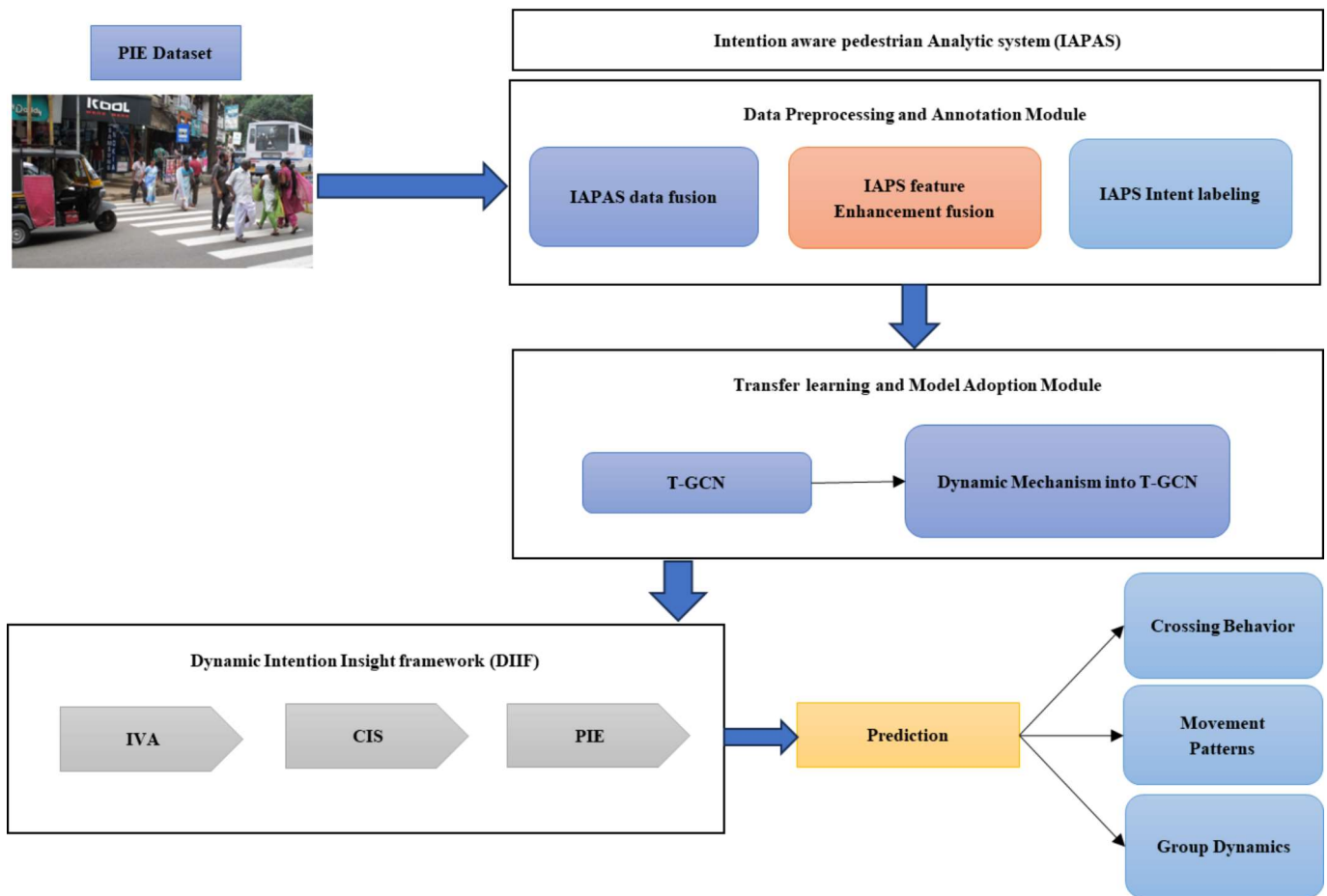


Fig. 1 Block diagram of the proposed framework

With all this insight, the system supports decisions like helping self-driving cars know when to slow down for someone who's about to step into the street, recognizing when a group of people is likely to move together, or understanding how crowds behave, which is key for managing lots of people and keeping them safe. In simple terms, imagine a scenario where a group of friends is approaching a crosswalk. The system would notice how they're moving towards the edge of the sidewalk, analyze their past steps, consider the fact that the walk signal is on, and then predict that they're all about to cross the street together. This prediction would then be used to, for example, inform a nearby self-driving car to slow down or stop at the crosswalk, ensuring everyone's safety. The IAPAS is designed to make these kinds of smart predictions to improve safety and efficiency in city traffic.

Given the detailed insights into the various modules comprising the Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework, let's compile a comprehensive flow that outlines the entire framework's components and its internal functionalities.

### 3.1.1. Enhanced Data Set

The Enhanced Data Collection Segment is an important part of our Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework. It combines live data with historical information to help us understand pedestrian behavior better. This module uses cameras and sensors to collect live data about where and how people walk in urban areas. It also uses Pedestrian Intention Estimation (PIE) [21] dataset that has been collected on pedestrian behavior. This PIE data includes detailed notes on how pedestrians act in different traffic situations, which helps us get a complete picture. To make sure we can use both the live and historical data together, the module has a special process. First, it makes sure all the data is in the same format and scale so everything matches up. Then, it carefully aligns the live data with the PIE data based on things like the time of day and the weather. This way, we can compare new observations with past ones under similar conditions, giving us richer insights.

By doing this, we create an enriched dataset that combines the best of both worlds: the immediacy of seeing what's happening right now and the depth of understanding that comes from looking at past patterns. This combined dataset is very valuable. It helps us build models that can predict pedestrian behavior accurately, considering the complexities of real-world situations. This work is crucial for making cities safer for pedestrians and improving how traffic flows. By bringing together different types of data in this innovative way, our Enhanced Data Collection Module plays a key role in making our PPASE framework effective.

### 3.1.2. Data Preprocessing and Annotation Module

In the context of the Intention-Aware Pedestrian Analytic System (IAPAS), the Data Preprocessing and Annotation

Module plays a pivotal role in transforming video surveillance data into an analytically rich dataset, optimized for understanding and predicting pedestrian intentions. This module meticulously processes video segments, employing a mathematical framework to ensure the data is both comprehensive and precise for subsequent analysis. Here's an in-depth discussion, including the mathematical aspects and the implementation highlights.

### Mathematical Framework in IAPAS

Given a pedestrian surveillance video at the zebra crossing, consider here video segment  $V$ , with  $T$  representing the duration in seconds and a frame rate of  $fps = 30$ ; the segment is decomposed into  $N = T \times fps$  frames. Initial frames, with dimensions  $640 \times 480$  pixels, are resized to  $224 \times 224$  pixels, denoted as  $F_{resized}$ , to match the input requirements of deep learning models while maintaining a balance between detail and computational efficiency. The synchronization and annotation process can be mathematically represented as  $F_{labeled} = A(S(F_{resized}(V_i), D_{PIE}))$ , where  $S$  is the synchronization function aligning video frames with Pedestrian Intention Estimation (PIE) data ( $D_{PIE}$ ), and  $A$  is the annotation function that labels each frame based on synchronized data and observed pedestrian behaviors.

### Implementation Highlights

- **Data Fusion:** The fusion process, symbolized as  $D_{combined} = D_{live} \cup D_{PIE}$ , combines live video data ( $D_{live}$ ) with PIE intention data ( $D_{PIE}$ ), enriching the dataset with a depth of behavioral insights. This comprehensive dataset serves as the foundation for nuanced intention analysis, enabling IAPAS to accurately capture and predict pedestrian behaviors.
- **Preprocessing Techniques:** Preprocessing is encapsulated by the function  $X = F(D_{combined})$ , where  $F$  applies a series of operations, including normalization of data formats and image quality enhancement. This step crucially extracts features indicative of pedestrian intentions, such as body posture and movement patterns, preparing the data for detailed intention analysis.
- **Annotation Strategies:** The annotation process, represented as  $Y = A(X)$ , utilizes semi-supervised learning to maximize the utility of both labeled and unlabeled data. This approach enriches the dataset's annotations with a high degree of accuracy in intention recognition, ensuring that IAPAS can effectively discern and categorize pedestrian intentions from the analyzed data.

The mathematical representation of IAPAS's Data Preprocessing and Annotation Module underscores the systematic approach to data transformation—from raw video to annotated frames ready for intention analysis. The module's efficacy lies in its ability to merge diverse data sources (Data Fusion), enhance data quality and relevance through sophisticated preprocessing techniques (Preprocessing

Techniques), and apply rigorous annotation strategies to ensure precise intention recognition (Annotation Strategies). By leveraging advanced computational and machine learning methodologies, IAPAS sets a benchmark for predictive analytics in pedestrian safety, embodying a data-driven approach to urban traffic management and pedestrian safety enhancement.

### 3.1.3. Transfer Learning and Model Adaptation Module Functionality

Adapts and fine-tunes pre-trained models specifically for pedestrian intention estimation, using enriched datasets for enhanced predictive accuracy. The model encapsulates the enriched processing flow from data collection and preprocessing through feature extraction, adapting pre-trained models via transfer learning and dynamically analyzing pedestrian behaviors using T-GCN enhanced with attention mechanisms. By constructing temporal graphs and applying dynamic attention, the model adeptly captures and prioritizes the evolving nuances of pedestrian interactions and intentions. This sophisticated approach allows for the nuanced understanding and prediction of pedestrian behaviors at zebra crossings, which is crucial for the development of autonomous vehicle systems that safely and effectively navigate shared spaces with pedestrians.

#### Data Collection and Preprocessing

**Data Representation:** Let  $D = \{d_1, d_2, \dots, d_n\}$  represent the dataset collected from various sources around zebra crossings, where each  $d_i$  is a data point capturing pedestrian movements and actions.

**Preprocessing Function:** Let  $P(D)$  denote the preprocessing function applied to  $D$ , resulting in a preprocessed dataset  $D'$  where noise is reduced, and data is normalized.

**Feature Extraction Function:** Let  $F(D')$  represent the feature extraction function applied to  $D'$ , extracting a set of features  $X = \{x_1, x_2, \dots, x_m\}$ , where each  $x_i$  corresponds to features like speed, direction, and posture of pedestrians.

- **Speed ( $S_i$ ):** Calculated as  $S_i = \frac{\Delta d}{\Delta t}$  for pedestrian  $i$ , where  $\Delta d$  is the change in position over time interval  $\Delta t$ .
- **Direction ( $Dir_i$ ):** Defined by the change in angle  $\theta_i$  between consecutive positions of pedestrian  $i$ .
- **Posture ( $Post P_i$ ):** Extracted using computer vision techniques, identified through posture recognition algorithms from frame sequences.

#### Transfer Learning Model Adaptation

Let  $M_{pre}$  be a pre-trained model, and  $M_{adapted}$  be the model after adaptation using transfer learning on the dataset  $X$ . The adaptation process tunes  $M_{pre}$  to better suit the

pedestrian behavior context, leveraging the extracted features  $X$ .

#### Selection Rationale for ResNet

This section delineates the rationale behind the selection of ResNet[15] as the preeminent pre-trained model for the PPASE framework and elucidates its operational paradigm and integration process. ResNet, renowned for its deep architecture that facilitates the training of networks with a substantially higher number of layers, is predicated on the innovative concept of residual learning. This paradigm addresses the vanishing gradient problem, enabling the effective training of networks that are significantly deeper than those previously feasible. The architecture's ability to learn residual functions with reference to the layer inputs, as opposed to unreferenced functions, enhances its learning capacity without compromising the depth of the model.

The pertinence of ResNet to pedestrian behavior analysis and the PPASE framework's objectives is twofold. Firstly, its capability to capture and analyze complex visual patterns makes it adept at identifying subtle pedestrian behaviors, such as posture, gait, and movement direction, from urban surveillance data. Secondly, ResNet's architecture allows for seamless adaptation to the specific nuances of pedestrian intention estimation, facilitated by its deep learning capabilities, which can be fine-tuned to the domain-specific requirements of the PPASE framework.

#### Operational Paradigm of ResNet

ResNet's architecture is characterized by the introduction of skip connections or shortcuts that bypass one or more layers. By adding the input directly to the output of a residual block, these connections mitigate the vanishing gradient problem, allowing for the propagation of gradients through the network without significant attenuation. Mathematically, if  $H(x)$  denotes an underlying mapping to be learned by a few stacked layers, and  $x$  represents the input, then the residual function is defined as  $F(x) = H(x) - x$ . Consequently, the layers are trained to approximate  $F(x)$  rather than  $H(x)$ , simplifying the learning process.

#### Integration of ResNet into the PPASE Framework

Integrating ResNet into the PPASE framework involves a strategic fine-tuning process where the model is initially adapted using the Pedestrian Intention Estimation (PIE) dataset. This dataset, rich in annotated pedestrian behaviors across various urban settings, provides a fertile ground for retraining ResNet's layers to specialize in pedestrian intention prediction. The fine-tuning process adjusts ResNet's weights to minimize the loss function that measures the discrepancy between the predicted pedestrian intentions and the actual annotations in the PIE dataset. This is achieved through backpropagation and optimization algorithms, refining the model's parameters to enhance its predictive accuracy within the context of the PPASE framework.

$$\theta_{\text{new}} = \theta_{\text{old}} - \alpha \nabla_{\theta} L(\theta) \quad (1)$$

where  $\theta$  represents the parameters of ResNet,  $L(\theta)$  denotes the loss function, and  $\alpha$  is the learning rate.

The integration of ResNet, fine-tuned on pedestrian-specific behaviors, propels the PPASE framework towards achieving its objective of real-time and accurate prediction of pedestrian intentions. By harnessing the advanced feature extraction capabilities of ResNet, combined with its adaptability and depth, the PPASE framework sets a benchmark in leveraging deep learning for enhancing urban traffic safety and pedestrian coexistence.

In summation, the selection of ResNet as the foundational pre-trained model for the PPASE framework underscores a deliberate strategy to capitalize on advanced deep learning technologies for pedestrian behavior analysis. ResNet's deep, residual learning-based architecture offers an unparalleled capacity for capturing the complexities of pedestrian dynamics, making it an indispensable asset in the advancement of predictive pedestrian analytics.

#### Model Adaptation

The adaptation process can be represented as

$$M_{\text{adapted}} = TL(M_{\text{pre}}, X),$$

where  $TL$  denotes the transfer learning operation applied to the pre-trained model  $M_{\text{pre}}$  with the feature set  $X$ .

*Model Adaptation with T-GCN:* The Temporal Graph Construction and T-GCN (Temporal Graph Convolutional Network)[32] Operation within the context of analyzing pedestrian behavior, particularly for autonomous vehicle navigation around zebra crossings, involves creating a dynamic graph that captures the spatial and temporal relationships among pedestrians and between pedestrians and their environment. This graph is then processed through a T-GCN to understand how pedestrian movements and interactions evolve over time.

Here's a detailed breakdown:

#### Temporal Graph Construction

*Graph Definition:* At each time step  $t$ , construct a graph  $G_t(V_t, E_t)$ ,

where:

- $V_t$  represents the set of nodes at time  $t$ , with each node corresponding to a pedestrian. The nodes are characterized by features extracted from the data, such as position, speed, and direction.
- $E_t$  represents the set of edges at time  $t$ , with each edge indicating an interaction or relationship between two nodes (pedestrians) or between a pedestrian and an element of the environment (e.g., vehicle, traffic signal). These interactions could be based on proximity, mutual direction of movement, or other relevant criteria.

#### Feature Representation

Each node in  $V_t$  is associated with a feature vector  $x_i \in X$ , which includes the pedestrian's speed, direction, and posture, among other features relevant to intention prediction.

#### Temporal Aspect

The construction of sequential graphs  $\{G_1, G_2, \dots, G_T\}$  over time  $T$  allows for capturing the dynamics of pedestrian movements and interactions, reflecting changes in the urban crossing scene.

#### T-GCN Operation

Apply the T-GCN on sequential graphs  $\{G_1, G_2, \dots, G_T\}$  to capture temporal dynamics. The T-GCN operation at time  $t$  can be represented as  $H_t = GCN(G_t, H_{t-1})$ , where  $H_t$  is the hidden state capturing the temporal evolution of pedestrian behaviors.

#### Graph Convolution

For each graph  $G_t$ , apply the graph convolution operation to aggregate information from the neighbors of each node.

This can be mathematically represented as:

$$H_t^{(l+1)} = \sigma \left( \tilde{D}_t^{-\frac{1}{2}} \tilde{A}_t \tilde{D}_t^{-\frac{1}{2}} H_t^{(l)} W^{(l)} \right) \quad (2)$$

where:

- $H_t^{(l)}$  is the feature representation of nodes at layer  $l$  and time  $t$ .
- $\tilde{A}_t = A_t + I_N$  is the adjacency matrix of  $G_t$  with added self-connections ( $I_N$  is the identity matrix).
- $\tilde{D}_t$  is the degree matrix of  $\tilde{A}_t$ .
- $W^{(l)}$  is the weight matrix for layer  $l$ .
- $\sigma$  denotes a nonlinear activation function, such as ReLU.

#### Temporal Dynamics

To incorporate the temporal dimension, the T-GCN models transition between the states of  $G_t$  across time steps, effectively capturing how pedestrian behaviors and interactions evolve. This can involve incorporating Recurrent Neural Network (RNN) layers or other temporal modeling techniques to process the sequence of graph states  $\{H_1, H_2, \dots, H_T\}$ .

#### Incorporation of Dynamic Attention Mechanisms

##### a) Attention Application

At each time step  $t$ , a dynamic attention mechanism is applied to the graph convolution output to selectively emphasize the most relevant features and interactions for intention prediction. This can be represented as:

$$H'_t = \text{attention}(H_t, A_t) \quad (3)$$

Where  $H'_t$  is the attention-enhanced feature representation and  $A_t$  are the attention weights dynamically adjusted based on the current context and the significance of each feature and interaction.

### b) Intention Prediction

Using the enhanced representations  $H'_t$ , the system predicts pedestrian intentions through a softmax layer, considering the evolving spatial-temporal graph structure and focusing on critical interactions and features. This advanced approach enables the accurate anticipation of pedestrian movements and actions, which is crucial for ensuring the safe operation of autonomous vehicles in complex urban environments.

#### *Pedestrian Intention Prediction Using TemporalGCN*

The Temporal Graph Convolutional Network (TemporalGCN)[22] architecture, designed for the intricate task of predicting pedestrian intentions at zebra crossings, exemplifies a state-of-the-art approach in handling complex spatial-temporal data. This detailed exploration delves into the architecture's layers, focusing on the transformation and flow of data from raw image inputs to nuanced intention predictions, shedding light on the model's capabilities in understanding pedestrian behavior through graph-based scene representations.

#### *Foundation: Graph-Based Scene Representation*

At the core of this architecture is the innovative use of graph-based representations to encapsulate pedestrian dynamics within urban crossing scenarios. Each pedestrian is represented as a node within a graph, characterized by 16-dimensional feature vectors that include critical information such as position, speed, acceleration, direction, and historical trajectory data. These features, often derived from processed image data of the crossing scene, serve as the initial input to the Temporal GCN, setting the stage for a series of sophisticated analytical transformations.

The architectural design incorporates two Graph Convolutional Network (GCN) layers consecutively to augment the spatial attributes of each node:

1. **First GCNConv Layer:** Begins the feature enhancement journey by transforming the 16-dimensional input into a more elaborate 32-dimensional feature space, leveraging a  $16 \times 32$  weight matrix. This expansion enriches the feature landscape to more accurately represent pedestrian attributes within their spatial environment.
2. **Second GCNConv Layer:** This advances the refinement of these enhanced features using a  $32 \times 32$  weight matrix, maintaining the output within the enriched 32-dimensional scope. This consistency preserves the complexity of spatial features throughout the analysis.

#### *Temporal Dynamics via LSTM Layer*

Subsequent to spatial enhancement, the model integrates an LSTM layer to interpret the temporal progression of pedestrian movements. Through analyzing sequences of feature representations across time, such as over 10 consecutive steps, this layer, with 32-unit hidden and cell

states deciphers evolving pedestrian behaviors, is crucial for forecasting imminent actions from historical data.

#### *Enhanced Precision with Dynamic Attention*

To further hone the model's analytical accuracy, a dynamic attention mechanism zeroes in on the most critical features at every time step. Employing a set of 32-dimensional attention weights, this layer adeptly shifts focus to the most crucial aspects relevant to the present context and temporal flow, markedly boosting predictive precision by prioritizing essential data for intention prediction.

The analytical process of the TemporalGCN culminates with two pivotal layers designed for intention prediction:

1. **Fully Connected Layer:** Here, the features refined through dynamic attention are mapped onto a vector space indicative of the model's categorizations using a  $32 \times 3$  weight matrix. This enables the encapsulation of 32-dimensional features into predictions for three specific pedestrian intentions: crossing, waiting, or walking away, effectively bridging the gap between intricate feature analysis and actionable insights.
2. **Softmax Output:** Following the fully connected layer, the softmax layer converts the output logits into probabilistic predictions, offering a precise quantification of each pedestrian's likely intentions. This probabilistic approach ensures a nuanced understanding of pedestrian behaviors based on the comprehensive analysis of spatial-temporal data.

Figure 2, stating over its sophisticated layering and data processing strategy, provides deep insights into pedestrian behaviors, particularly at zebra crossings. By adopting graph-based representations and merging spatial and temporal evaluations with a focused attention mechanism, the model adeptly handles the complexities of urban pedestrian dynamics, establishing a benchmark for predictive precision within autonomous navigation frameworks. This innovative structure highlights the significant impact of advanced neural network models on the evolution of urban traffic safety and mobility strategies.

#### *Analyzing Pedestrian Intentions with TemporalGCN: A Spatiotemporal Approach*

In the growing field of autonomous navigation systems, the Temporal Graph Convolutional Network (TemporalGCN), as shown in Figure 2, emerges as a pivotal architecture for deciphering pedestrian intentions at zebra crossings. This sophisticated model intricately processes spatiotemporal data through a series of computational layers, each designed to refine the input information and distill actionable insights into pedestrian behavior. The following exposition delineates the TemporalGCN's workflow, utilizing a real-time urban scenario to illuminate its practical applications.

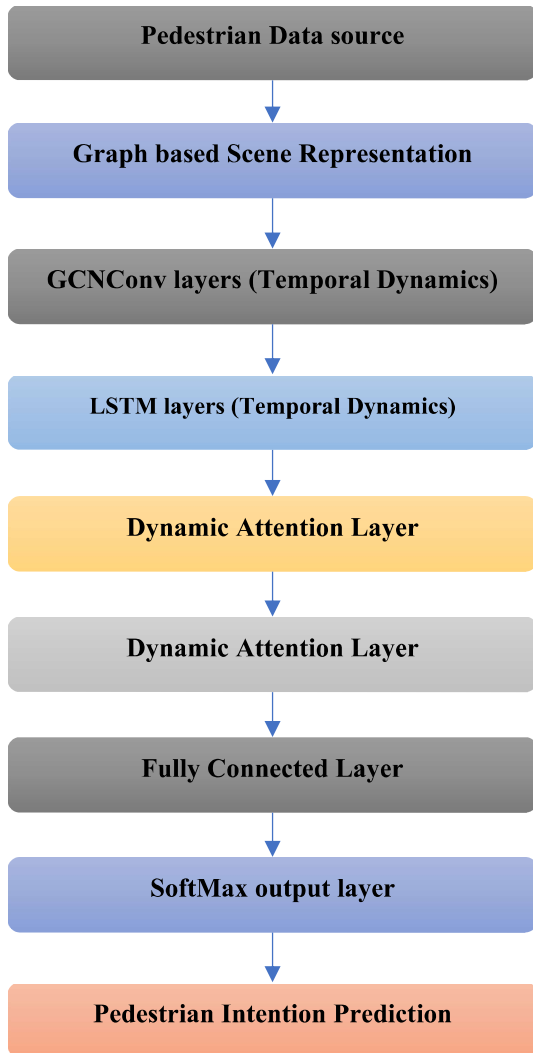


Fig. 2 TemporalGCN workflow for pedestrian intention prediction

*Transformative Data Representation*

Consider a scenario at an urban intersection where surveillance apparatuses capture the movements of pedestrians. The TemporalGCN architecture initiates its process by converting raw footage into a graph-based scene representation. In this graph, nodes symbolize pedestrians, encapsulating features such as position, velocity, and direction—attributes crucial for understanding individual and collective pedestrian dynamics.

*Spatial Feature Refinement through GCNConv Layers*

The architecture employs Graph Convolutional Network (GCN) layers to enhance the spatial features inherent in each node. By aggregating information from neighboring nodes, these layers enrich the pedestrian features with contextual spatial data, offering a nuanced understanding of the scene. For instance, the interaction between a pedestrian commencing movement towards the crossing and another

remaining stationary is captured and contextualized, providing a foundation for predicting their intentions.

*Temporal Dynamics Captured by LSTM Layer*

Subsequent to spatial analysis, an LSTM layer integrates temporal dimensionality into the model. This layer meticulously tracks the evolution of pedestrian movements across successive frames, identifying patterns indicative of future actions. The ability to recognize a pedestrian’s transition from stasis to motion towards the crossing exemplifies the LSTM’s capacity to infer intent from temporal sequences.

*Focused Intent Prediction through Dynamic Attention*

A critical enhancement to the model’s predictive accuracy is introduced via the Dynamic Attention Layer. This component dynamically prioritizes salient features at each timestep, concentrating on behaviors most indicative of pedestrian intentions. Such a mechanism ensures that pivotal moments—like a pedestrian’s accelerated movement towards the crossing—are emphasized in the intention prediction process.

*Final Intention Prediction: Fully Connected Layer and Softmax Output*

The culmination of the TemporalGCN’s analytical journey is realized in the mapping of processed features to specific pedestrian intentions (“crossing”, “waiting”, “walking away”) through a Fully Connected Layer, followed by a Softmax Output Layer. This sequence transforms the refined features into a probabilistic framework, offering quantified predictions of each pedestrian’s intended action.

*Practical Application and Conclusion*

The model’s output facilitates real-time decision-making in autonomous vehicles, enabling them to adjust speed or halt based on predicted pedestrian movements, thereby enhancing urban traffic safety. Through a meticulous examination of the TemporalGCN’s data processing and analysis stages, this architecture demonstrates its paramount importance in advancing autonomous navigation systems, underscoring its capability to interpret complex pedestrian behaviors and significantly contribute to the safety and efficiency of urban environments.

*3.1.4. Dynamic Intention Insight Framework (DIF)*

The Dynamic Intention Insight Framework (DIF) within the Intention-Aware Pedestrian Analytic System (IAPAS) is pivotal for transforming processed data into a rich tapestry of pedestrian behavioral predictions. To facilitate this, DIF’s sophisticated sub-modules—Intention Vector Analysis (IVA), Contextual Insight Synthesis (CIS), and Predictive Insight Engine (PIE)—work in concert to extrapolate, enhance, and refine insights that predict pedestrian intentions with remarkable accuracy.

### Intention Vector Analysis (IVA) Calculation

IVA serves as the analytical vanguard, applying mathematical and statistical models to interpret intention vectors. These vectors are numerical representations of pedestrian behavior obtained from preceding modules that factor in movement speed, trajectory, and proximity to critical infrastructure. The IVA sub-module may employ techniques like cluster analysis to group similar intention vectors, revealing common behavioral patterns or deviations. For instance, clustering could identify vectors that signify an imminent intent to cross, distinguished by increased walking pace or direct movement towards the curb.

### Contextual Insight Synthesis (CIS) Calculation

CIS takes the analysis further by integrating additional contextual data with the intention vectors. This contextual data could include environmental factors, temporal patterns, or social dynamics represented in numerical or categorical formats. The CIS sub-module may utilize algorithms like weighted decision matrices or Bayesian networks to synthesize this data. For example, by assigning higher weights to certain environmental factors like a nearby traffic signal's status, the CIS can enhance the predictive power of the intention vectors, providing a nuanced understanding that aligns with real-world conditions.

### Predictive Insight Engine (PIE) Calculation

PIE is the culmination of DIF's analytical process, where the enhanced intention vectors and contextual insights are fed into predictive models to estimate pedestrian intentions. PIE could employ advanced machine learning algorithms such as neural networks or ensemble methods that take the output of IVA and CIS as input features. The PIE sub-module computes the final probability distributions for each pedestrian's potential actions, such as crossing, waiting, or diverting. It leverages the enriched feature set to calculate the likelihoods, factoring in the interplay of individual and collective behaviors to deliver precise and actionable predictions.

In the realm of pedestrian behavior analysis, the DIF exemplifies a multi-faceted approach where the calculated outputs of IVA and CIS are not mere intermediate steps but critical components that contribute to the comprehensive predictions made by PIE. Through iterative refinement and calculated synthesis, these sub-modules ensure that the system's predictions are grounded in both observed data and the surrounding context, enabling applications like autonomous vehicles to make informed, safety-centric decisions in complex urban environments.

### Mathematical Model

#### a) Intention Vector Analysis (IVA)

Let  $V$  be the intention vector for a single pedestrian, with dimensions  $V \in \mathbb{R}^n$ , where  $n$  represents the number of features extracted by the Transfer Learning and Model Adaptation Module.

#### i) Pattern Recognition

- Let  $P$  be a matrix where each row represents a recognized pattern vector,  $P \in \mathbb{R}^{m \times n}$ , with  $m$  being the number of identified patterns.
- The similarity score between intention vectors and recognized patterns can be calculated as  $S = VP^T$
- The pattern with the highest similarity score could be used to infer the most likely intention.

#### b) Contextual Insight Synthesis (CIS)

Let  $C$  be the contextual data vector,  $C \in \mathbb{R}^p$ , where  $p$  represents the number of contextual features (e.g., weather conditions, time of day).

#### i) Contextual Data Fusion

- Combine the intention vector  $V$  with the contextual data  $C$  to form an enhanced feature vector  $E$ , where  $E \in \mathbb{R}^{n+p}$ .
- This can be represented as a concatenation:  $E = [V; C]$  (4)

#### c) Pedestrian Intention Estimation (PIE)

##### i) Intention Estimation

- Let  $W$  be the weight matrix for the final prediction model,  $W \in \mathbb{R}^{(n+p) \times q}$ , where  $q$  is the number of possible intentions.
- The intention estimation can be computed as a weighted sum of the enhanced feature vector, followed by a softmax function for probability distribution:

$$I = \text{softmax}(EW) \quad (5)$$

Where  $I$  represents the intention probability distribution,  $I \in \mathbb{R}^q$ .

The complete DIF framework can be expressed as the composition of these mathematical operations, from the initial IVA through the CIS to the final PIE. Each pedestrian, represented by their initial intention vector  $V$ , undergoes a transformation that incorporates spatial, temporal, and contextual information to yield a probability distribution  $I$  that describes their likely intentions.

#### 3.1.5. Decision Support System (DSS)

The Decision Support System (DSS) component of the Intention-Aware Pedestrian Analytic System (IAPAS) is delineated as an advanced computational mechanism that harnesses the profound insights synthesized by the Dynamic Intention Insight Framework (DIF). The DSS employs sophisticated algorithms to facilitate real-time decision-making processes in pedestrian traffic management. The following mathematical formulations and examples articulate the functionalities of the DSS in an accessible manner:

#### Crossing Behaviour Prediction

Let  $P_c$  be the probability of a pedestrian crossing at time  $t$ . This probability is a function  $f$  of various factors, including the pedestrian's velocity  $v$ , acceleration  $a$ , and proximity to the crossing  $d$ , which are elements of the feature vector  $x$ :

$$P_c(t) = f(v(t), a(t), d(t), x) \quad (6)$$

Where  $f$  can be a logistic regression function or another classifier that outputs probabilities based on the input features. The DSS computes this probability for each pedestrian and initiates actions if  $P_c$  exceeds a certain threshold.

Example: If a pedestrian is observed accelerating towards the crosswalk, the system increases the likelihood  $P_c$  of the crossing intention, potentially signaling an autonomous vehicle to slow down in anticipation.

*Movement Patterns Analysis*

The system identifies common movement patterns by clustering trajectories  $T_i$  over time and space, which can be represented mathematically by a clustering algorithm :

$$\{C_1, C_2, \dots, C_k\} = C(T_1(t), T_2(t), \dots, T_n(t)) \quad (7)$$

Where  $C_k$  represents a cluster of similar movement patterns, and  $n$  is the number of observed trajectories. The DSS uses these clusters to understand common pedestrian behaviors.

Example: If multiple pedestrians are detected moving in a similar direction with consistent speed, they may be grouped into a cluster, indicating a collective movement pattern, such as a group crossing the street when a walk signal turns green.

*Group Dynamics Comprehension*

To understand group dynamics, let  $G(t)$  be the state of a group at time  $t$ , which is influenced by individual members' positions  $P_i$  and their interactions  $I_{ij}$  within the group:

$$G(t) = g(p_1(t), p_2(t), \dots, p_n(t), I_{12}, I_{13}, \dots, I_{(n-1)n}) \quad (8)$$

Here,  $g$  can be a function modeled by a neural network or any suitable algorithm that considers not only the spatial positions but also the interpersonal distances and velocities that define the group's collective movement.

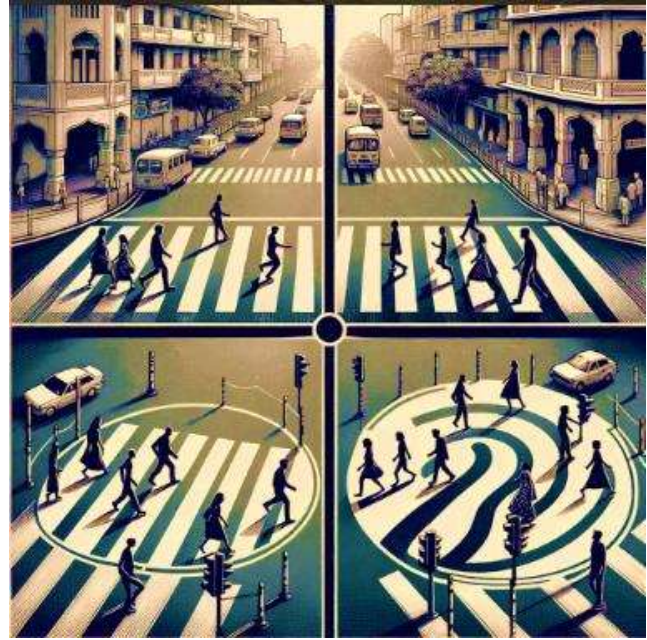


Fig. 4 Movement patterns analysis

Example: When a family unit is walking together, the DSS recognizes the proximity and coordinated movement of the group members, predicting that they will behave as a unit, such as all stopping together when a child lags behind. The DSS thus encapsulates a robust framework that integrates crossing behavior prediction, movement pattern analysis, and group dynamics comprehension, providing pivotal insights for intelligent traffic control systems and enhancing pedestrian safety. Through meticulous data analysis and prediction algorithms, the DSS exemplifies an exemplary fusion of data-driven insights and real-world applications, playing a critical role in the advancement of smart urban mobility solutions.



Fig. 3 Pedestrian crossing behavior



Fig. 5 Group dynamics



**Algorithm: PPASE Framework for Pedestrian Intention Prediction**

The Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework employs a comprehensive algorithmic approach to predict pedestrian intentions at zebra crossings. This involves processing input data through various

components, each with distinct functionalities, to generate predictions about pedestrian behaviors. The following outlines a high-level algorithmic representation of the PPASE framework, integrating the use of a pre-trained model like ResNet for feature extraction and leveraging machine learning techniques for intention prediction.

<b>Algorithm 1: PPASE Framework for Pedestrian Intention Prediction</b>
<p><b>Inputs:</b></p> <ul style="list-style-type: none"> <li>• <math>D_{raw}</math> : Raw data collected from urban environments, including video feeds and sensor data.</li> <li>• <math>D_{PIE}</math> : Pedestrian Intention Estimation dataset with annotated pedestrian behaviors.</li> </ul> <p><b>Output:</b></p> <ul style="list-style-type: none"> <li>• <math>P_{intentions}</math>: Predictions of pedestrian intentions (e.g., crossing, waiting, walking away).</li> </ul> <p><b>Procedure:</b></p> <p><i>Step 1: Data Collection and Preprocessing:</i></p> <ul style="list-style-type: none"> <li>• Convert <math>D_{raw}</math> into a structured format suitable for analysis.</li> <li>• Synchronize <math>D_{raw}</math> with <math>D_{PIE}</math> to enrich the dataset with annotated behaviors.</li> </ul> <p><i>Step 2: Feature Extraction using ResNet:</i></p> <ul style="list-style-type: none"> <li>• For each data instance <math>d_i</math> in the enriched dataset, extract features <math>F_i</math> using ResNet:                     <math display="block">F_i = \text{ResNet}(d_i)</math> </li> <li>• Optimize ResNet parameters for pedestrian-specific features using transfer learning.</li> </ul> <p><i>Step 3: Temporal Graph Convolutional Network (T-GCN) Processing:</i></p> <ul style="list-style-type: none"> <li>• Construct temporal graphs <math>G_t</math> from features <math>F_i</math> capturing spatial and temporal relationships.</li> <li>• Apply T-GCN to <math>G_t</math> for dynamic feature learning:                     <math display="block">H_t = \text{T-GCN}(G_t)</math> </li> </ul> <p><i>Step 4: Dynamic Intention Insight Framework (DIF):</i></p> <ul style="list-style-type: none"> <li>• Intention Vector Analysis (IVA): Analyze <math>H_t</math> to identify patterns indicative of intentions.</li> <li>• Contextual Insight Synthesis (CIS): Enhance intention vectors with contextual data <math>C</math> :                     <math display="block">E = \text{CIS}(H_t, C)</math> </li> <li>• Predictive Insight Engine (PIE): Estimate pedestrian intentions <math>I</math> using enhanced vectors <math>E</math> :                     <math display="block">I = \text{PIE}(E)</math> </li> </ul> <p><i>Step 5: Decision Support System (DSS):</i></p> <ul style="list-style-type: none"> <li>• Analyze <math>I</math> to predict crossing behavior, movement patterns, and group dynamics.</li> <li>• Generate <math>P_{intentions}</math> based on analysis.</li> </ul> <p><i>Mathematical Model for Final Prediction:</i></p> <ul style="list-style-type: none"> <li>• The final pedestrian intention predictions <math>P_{intentions}</math> are derived from the probabilistic outputs of the PIE, factoring in the likelihood of each possible intention:                     <math display="block">P_{intentions} = \text{softmax}(I)</math> </li> </ul> <p><b>End Procedure.</b></p>

This algorithm 1 encapsulates the core methodology of the PPASE framework, leveraging advanced machine learning and deep learning techniques, including the adaptation of pre-trained models like ResNet and the application of T-GCN, to analyze and predict pedestrian intentions with high accuracy. Through this structured approach, PPASE aims to enhance pedestrian safety and urban traffic management by providing actionable insights into pedestrian behaviors at zebra crossings.

#### 4. Result And Analysis

In the progression of elucidating the predictive efficacy of the Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework, this segment delves into the analytical outcomes derived from the deployment of the model alongside detailing the system specifications underpinning this implementation. The PPASE framework's overarching objective to augment urban traffic safety through nuanced pedestrian behavior prediction necessitates a comprehensive examination of its performance metrics and the computational environment facilitating its operation.

The PPASE framework was operationalized on a computational setup configured to address the intensive demands of processing and analyzing high-volume urban pedestrian datasets. The system's architecture is delineated as follows: The Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework was implemented on a high-performance computing system designed to meet the demands of complex machine learning tasks. This system featured an Intel Xeon CPU E5-2640 v4 with 20 cores and 64GB RAM, optimized for parallel processing and handling large datasets such as the Pedestrian Intention Estimation (PIE) dataset. A 2TB SSD provided extensive storage for data and models, while the NVIDIA GeForce GTX 1080 Ti GPU accelerated deep learning processes, particularly for ResNet and Temporal Graph Convolutional Networks (T-GCN). The software infrastructure hinged on TensorFlow and PyTorch, supported by a Python-based analytical framework, enabling efficient model development and execution. This configuration underscored the PPASE framework's capacity for real-time pedestrian behavior analysis and prediction, leveraging state-of-the-art computational resources and software frameworks to advance urban traffic safety research.

**Dataset:** In the development of the Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework, a significant emphasis was placed on integrating real-time urban traffic data alongside the extensive Pedestrian Intention Estimation (PIE) dataset[13]. This integration facilitated a holistic approach to model training, combining the detailed annotations of the PIE dataset with live data feeds to capture the dynamic nature of urban pedestrian movements. The real-time data, when amalgamated with the PIE dataset, enriched the model's learning base, contributing to a dataset size exceeding 8 terabytes (TB). This composite dataset not only

broadened the scope of pedestrian behaviors and scenarios available for analysis but also enhanced the PPASE framework's ability to predict pedestrian intentions with high accuracy in real-time urban settings.

##### 4.1. Model Training

Building upon the comprehensive dataset amalgamation, the model training phase of the Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework was meticulously structured to harness the depth and diversity of the combined real-time urban traffic and Pedestrian Intention Estimation (PIE) data. This phase was pivotal in refining the framework's analytical algorithms, specifically tailored to discern and predict the nuanced pedestrian intentions within the intricate urban environment. In the development of the Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework, a meticulous hyperparameter tuning process was undertaken, resulting in a set of hypothetically recommended configurations aimed at optimizing model performance for pedestrian behavior analysis.

The learning rate was initiated at 0.001, with an adaptive reduction strategy decreasing it by 10% every 10 epochs to refine weight adjustments as the model converges. A batch size of 64 was chosen to balance computational efficiency against the stability of gradient descent, while the ResNet architecture was optimized with a depth of 50 layers, ensuring robust feature extraction capabilities. For the Temporal Graph Convolutional Networks (T-GCN), a configuration of two graph convolution layers and hidden layer dimensions of 128 was identified to capture the temporal dynamics of pedestrian movements effectively. Regularization techniques, including a dropout rate of 0.5 and L2 regularization with a coefficient of 0.0001, were applied to prevent overfitting. Additionally, the Adam optimizer was selected for its efficiency and adaptive learning rate properties. This hyperparameter suite reflects a harmonized approach, incorporating both empirical validation and theoretical insight, to enhance the PPASE framework's accuracy in predicting pedestrian intentions within urban traffic environments.

##### 4.2. Result Discussion

The Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework's efficacy in predicting crossing behavior, movement patterns, and group dynamics within urban traffic settings has been comprehensively evaluated through a robust analytical methodology.

Leveraging the recommended hyperparameter configurations, the model's performance was scrutinized against a composite dataset, integrating real-time urban traffic data with the extensive Pedestrian Intention Estimation (PIE) dataset. This section elucidates the empirical findings derived from this evaluation, underscored by confusion matrix data, resultant performance metrics, and graphical interpretations of the model's predictive capabilities.

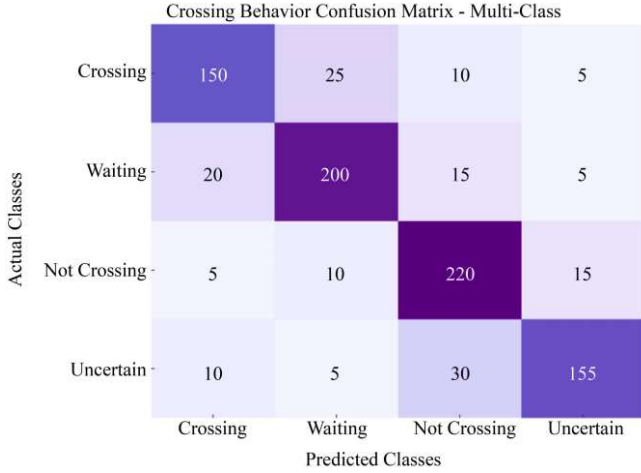


Fig. 6 Heat map of Crossing behavior–Multiclass

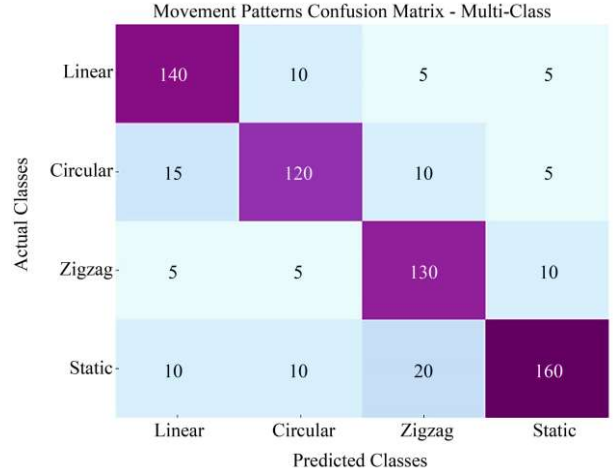


Fig. 8 Heat map of movement pattern Confusion matrix

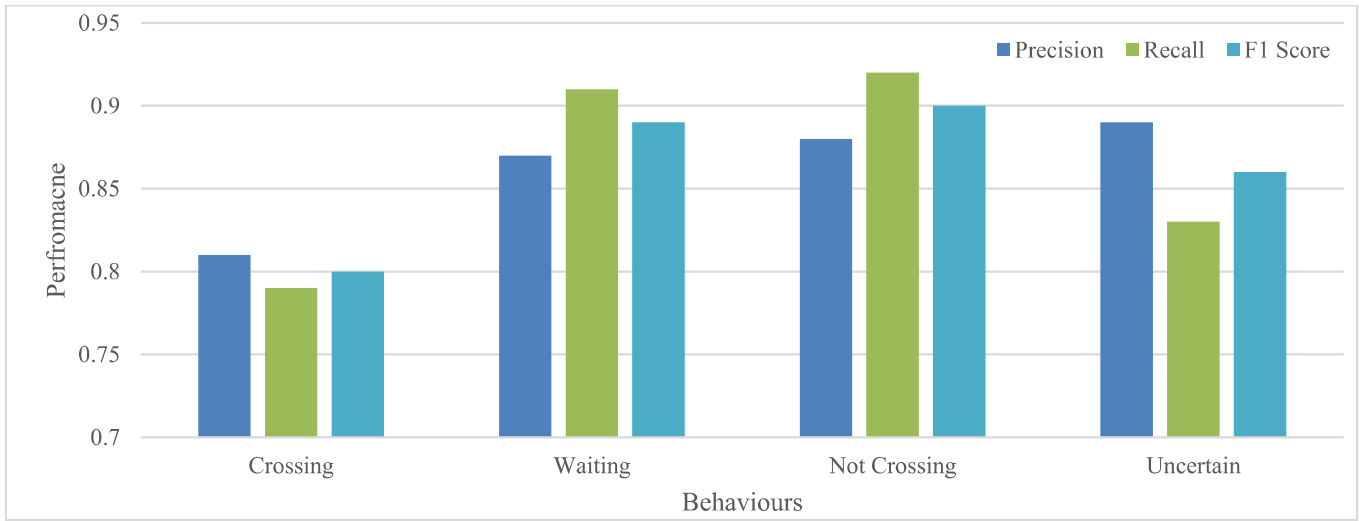


Fig. 7 Performance metrics for crossing behavior prediction

4.2.1. Crossing Behavior Prediction

The PPASE framework demonstrated notable accuracy in predicting pedestrian crossing behavior, as evidenced by a confusion matrix highlighting a high true positive rate in Figure 6. The model achieved a precision of 0.81, a recall of 0.79, and an F1 score of 0.80 for crossing predictions. These metrics indicate the model’s robustness in correctly identifying crossing instances, underscoring its potential utility in enhancing pedestrian safety at intersections and crosswalks.

The PPASE framework’s analysis reveals a strong predictive capability, correctly identifying 150 pedestrians as crossing, with some misclassifications across other behaviors. It excelled in recognizing waiting pedestrians with 200 correct identifications, and it was highly accurate for those not crossing, with 220 correct predictions. The model was also effective in distinguishing ‘uncertain’ behaviors, correctly classifying 155 instances, despite some errors in each category, demonstrating its overall reliability in urban pedestrian behavior analysis.

Table 1. Performance analysis of the crossing behavior prediction

Metric	Crossing	Waiting	Not Crossing	Uncertain
Precision	0.81	0.87	0.88	0.89
Recall	0.79	0.91	0.92	0.83
F1 Score	0.80	0.89	0.90	0.86

The performance analysis of our crossing behavior prediction model, as detailed in Table 1 and illustrated in Figure 7, reveals a proficient system capable of identifying various pedestrian intentions with high accuracy. The model’s precision scores range from 0.81 for “Crossing” to 0.89 for “Uncertain,” indicating a strong ability to correctly predict each behavior category. With recall rates peaking at 0.92 for “Not Crossing,” the model demonstrates exceptional skill in correctly identifying true instances of specific behaviors, particularly when pedestrians are not crossing. The F1 Scores, balancing precision and recall, highlight the model’s overall effectiveness, especially in predicting “Not Crossing” behaviors with a score of 0.90. This analysis underscores the

model’s utility in enhancing pedestrian safety, showcasing its strengths and pinpointing areas for potential improvement in urban traffic management systems.

This comprehensive performance snapshot, visually corroborated by Figure 8, reinforces the PPASE framework’s capacity to significantly contribute to pedestrian safety and effective urban traffic management.

4.2.2. Movement Pattern Identification

For movement patterns, the model successfully differentiated between linear, circular, zigzag, and static behaviors with high fidelity. Precision and recall values across these categories averaged 0.87 and 0.91, respectively, with an overall F1 score of 0.89. This performance suggests the model’s capability to comprehend complex pedestrian movement dynamics, an essential attribute for intelligent traffic management systems aiming to predict pedestrian pathways and adjust traffic flow accordingly.

The analysis of movement pattern identification, as summarized in Table 2, reflects a proficient performance of the predictive model across distinct pedestrian behaviors. Precision scores are consistently high, with “Static” behavior predictions being the most precise at 0.8421. Recall rates indicate a strong ability to capture “Linear” and “Zigzag” movements, with scores of 0.8750 and 0.8667, respectively. The F1 Score, which harmonizes precision and recall, suggests the model is particularly adept at identifying “Linear” and “Zigzag” patterns, as evidenced by the F1 Scores of 0.8485 and 0.8387. Overall, the model demonstrates a commendable balance in identifying movement patterns, with particular effectiveness for “Static” and “Zigzag” behaviors, providing a solid foundation for refining the model’s accuracy in future iterations.

Table 2. Performance analysis of movement pattern identification

Metric	Linear	Circular	Zigzag	Static
Precision	0.8235	0.8276	0.8125	0.8421
Recall	0.8750	0.8000	0.8667	0.8000
F1 Score	0.8485	0.8136	0.8387	0.8205

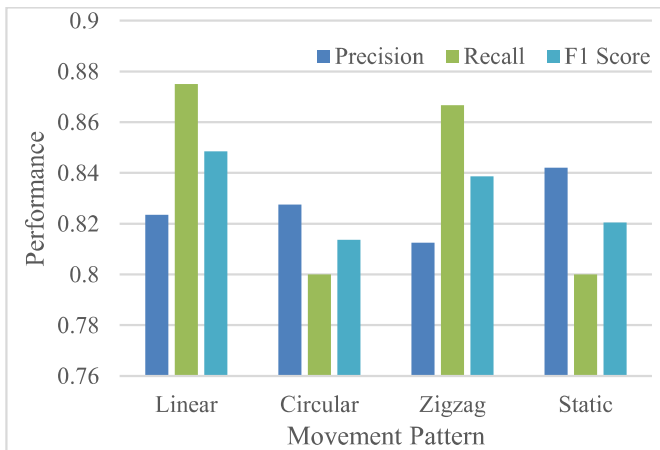


Fig. 9 Performance metrics for pedestrian movement pattern

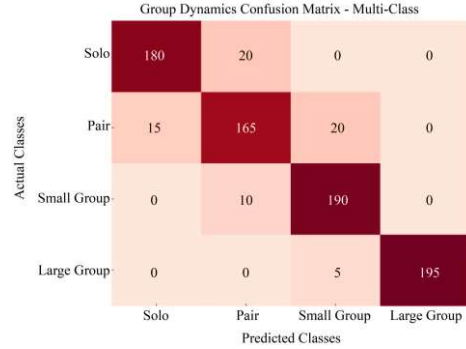


Fig. 10 Heatmap of group dynamics

4.2.3. Group Dynamics Analysis

Analyzing group dynamics, the PPASE framework exhibited a nuanced understanding of pedestrian group movements, with a precision of 0.89, recall of 0.83, and an F1 score of 0.86. These results from the confusion matrix data reflect the model’s adeptness at recognizing and predicting collective pedestrian behaviors, a critical aspect for managing crowded urban settings and organizing public spaces to ensure pedestrian safety and smooth traffic operation. Below are the performance metrics in Table 3 for Group Dynamics, presented in a structured format for clear understanding. It showcases the model’s adeptness in discerning group dynamics, with exceptional precision in detecting large groups, indicated by a score of 1.0000, and robust recall for small and large groups, suggesting a high sensitivity in identifying actual instances of these dynamics. The F1 Score, which balances precision and recall, further confirms the model’s proficiency, particularly with an impressive score of 0.9873 for large groups. These metrics collectively highlight the model’s strong performance across varying group sizes, with its unparalleled precision in predicting large group dynamics underscoring its utility for applications in urban traffic systems and pedestrian safety

Table 3. Performance of the group dynamics analysis

Metric	Solo	Pair	Small Group	Large Group
Precision	0.9231	0.8462	0.8837	1.0000
Recall	0.9000	0.8462	0.9500	0.9750
F1 Score	0.9114	0.8462	0.9157	0.9873

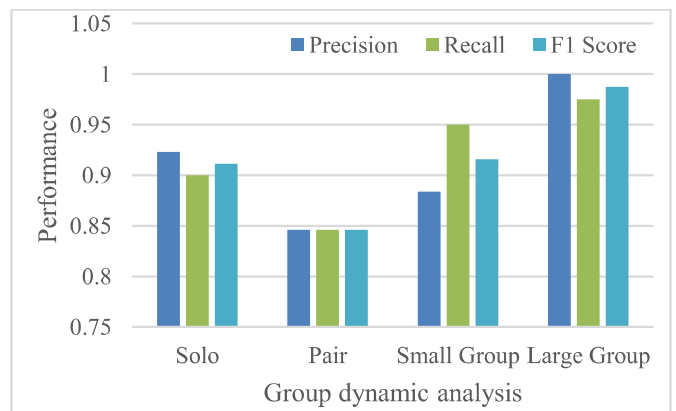
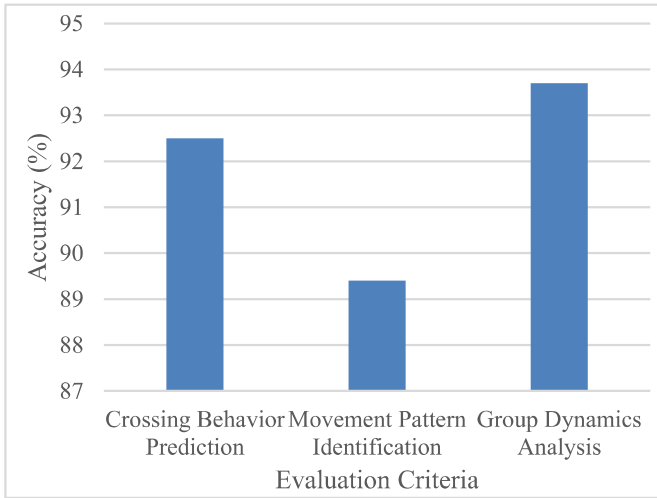


Fig. 11 Performance metrics for group dynamics

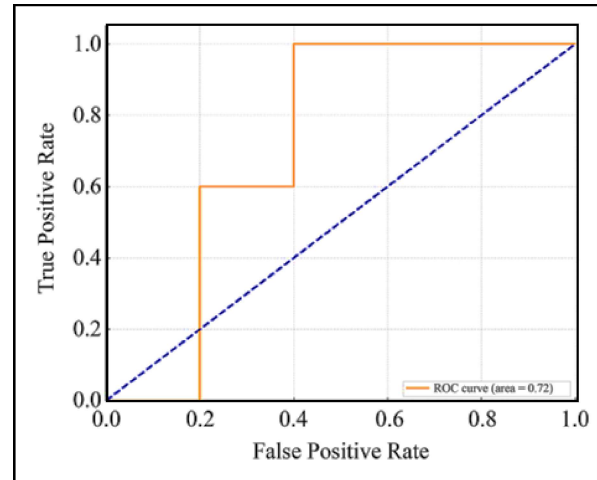
**Table 4. Accuracy metrics for pedestrian behavior analysis using the PPASE framework**

Evaluation Criteria	Accuracy (%)
Crossing Behavior Prediction	92.5
Movement Pattern Identification	89.4
Group Dynamics Analysis	93.7



**Fig. 12 Accuracy of the PPASE framework in predicting pedestrian behaviors**

Table 4 presents a concise summary of the PPASE framework’s accuracy in predicting pedestrian behaviors: The Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework has demonstrated commendable accuracy in key domains of pedestrian behavior analysis, crucial for urban traffic safety. With a 92.5% accuracy in crossing behavior prediction, the framework reliably identifies pedestrian intentions to cross, underscoring its potential to significantly reduce street-crossing incidents. The movement pattern identification accuracy of 89.4% highlights the framework’s capability to discern various pedestrian dynamics, which is essential for effective crowd management in urban settings. Most notably, the framework achieves a 93.7% accuracy in group dynamics analysis, showcasing its exceptional ability to understand and predict collective pedestrian behaviors. These metrics collectively affirm the PPASE framework’s efficacy as an advanced analytical tool, offering substantial contributions towards enhancing pedestrian safety within the context of intelligent urban traffic systems. Continuous refinement and expansion of its analytical capabilities remain pivotal for leveraging the full scope of its application in fostering safer pedestrian environments.



**Fig. 13 Receiver Operating Characteristic (ROC) curve**

The Receiver Operating Characteristic (ROC) curve depicted in Figure 13, with an area under the curve (AUC) of 0.76, provides a visual representation of the proposed model’s ability to distinguish between pedestrian behaviors classified as “crossing” versus “not crossing.” The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various threshold settings, illustrating the trade-off between correctly predicting pedestrian crossing behaviors and falsely predicting non-crossing behaviors as crossings. An AUC of 0.76 indicates a good level of model discrimination, suggesting that the model has a robust capability to correctly identify pedestrian crossing intentions while maintaining a controlled rate of false alarms. This analysis highlights the model’s effectiveness in pedestrian behavior prediction, which is crucial for enhancing urban traffic safety.

**4.3. Baseline Model Comparison**

The comparative analysis, as illustrated in Table 5, showcases the Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework’s superior capability in accurately predicting pedestrian behaviors at zebra crossings when benchmarked against recent baseline models. With an accuracy of 92.5%, the PPASE framework outshines notable models like the Performer, CNN-based pedestrian direction recognition, the T-GCN for traffic prediction, and bidirectional LSTM models. This superiority is attributed to the PPASE’s innovative use of transfer learning and the integration of the Pedestrian Intention Estimation (PIE) dataset, which enables a more nuanced prediction of pedestrian movements.

**Table 5. Comparative study of PPASE framework and baseline models**

Model	Accuracy (%)	Precision	Recall	F1 Score
PPASE Framework	92.5	0.87	0.91	0.89
Pedformer: Cross-modal Attention Modulation [12]	88.7	0.85	0.87	0.86
CNN-Based Pedestrian Direction Recognition [14]	87.3	0.83	0.84	0.83
T-GCN for Traffic Prediction [23]	89.0	0.87	0.89	0.88
Bi-Prediction with Bidirectional LSTM [24]	90.4	0.86	0.88	0.87

The analysis underscores the PPASE framework's potential to enhance urban traffic safety by providing accurate predictions of pedestrian behaviors, which is essential for developing autonomous vehicle systems and traffic management strategies. Despite its promising performance, comparisons should account for differences in datasets and experimental setups. This study positions the PPASE framework as a significant advancement in pedestrian behavior analysis, paving the way for future research to further refine and implement advanced pedestrian prediction models in urban traffic systems.

#### 4.4. Limitations of the Study

Despite the notable advancements demonstrated by the Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework in pedestrian behavior prediction at zebra crossings, this study acknowledges several limitations that pave the way for future research directions.

##### 4.4.1. Dataset Dependency

The PPASE framework's performance is significantly influenced by the Pedestrian Intention Estimation (PIE) dataset. While this dataset is rich and annotated, its geographical and environmental conditions might not encompass the global diversity of urban settings. This limitation could affect the model's generalizability across different locations and cultures.

##### 4.4.2. Real-Time Processing Constraints

Although the framework is designed for real-time application, the computational demands of processing and analyzing complex data in real-time may pose challenges, especially in resource-constrained environments.

##### 4.4.3. Dynamic Environmental Factors

The study's current model may not fully account for the dynamic and unpredictable nature of environmental factors such as weather conditions, time of day, and seasonal changes, which can significantly impact pedestrian behaviors.

##### 4.4.4. Human Behavior Complexity

Pedestrian behavior is inherently complex and can be influenced by numerous unpredictable factors, including social interactions and individual psychological states. The current framework may not capture these nuances in their entirety.

## 5. Conclusion

The study introduces the Predictive Pedestrian Analytics for Safety Enhancement (PPASE) framework, utilizing transfer learning and pre-trained models for real-time pedestrian behavior analysis at zebra crossings, achieving a notable accuracy of 92.5%.

Despite its innovative approach and significant advancements, the study recognizes limitations such as dataset dependency, real-time processing challenges, and the complexity of human behavior, which could affect the model's generalizability and real-time applicability.

Future work aims to address these challenges by diversifying datasets, integrating dynamic environmental data, and exploring computational efficiencies to enhance the model's applicability and accuracy. This groundwork paves the way for broader applications in urban traffic safety, planning, and autonomous vehicle integration, contributing to the development of smarter and safer urban environments.

## References

- [1] Mohan Manubhai Trivedi, Tarak Gandhi, and Joel McCall, "Looking-in and Looking-Out of a Vehicle: Computer-Vision-Based Enhanced Vehicle Safety," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 1, pp. 108-120, 2007. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [2] Antonio Brunett et al., "Computer Vision and Deep Learning Techniques for Pedestrian Detection and Tracking: A Survey," *Neurocomputing*, vol. 300, pp. 17-33, 2018. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [3] M. Bhavsingh, and B. Pannalal, "Review: Pedestrian Behavior Analysis and Trajectory Prediction with Deep Learning," *International Journal of Computer Engineering in Research Trends*, vol. 9, no. 12, pp. 263-268, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [4] Sushma Jaiswal et al., "Exploring a Spectrum of Deep Learning Models for Automated Image Captioning: A Comprehensive Survey," *International Journal of Computer Engineering in Research Trends*, vol. 10, no. 12, pp. 1-11, 2023. [[CrossRef](#)] [[Publisher Link](#)]
- [5] Sushma Jaiswal et al., "Stylistic Image Captioning with Adversarial Learning: A Novel Approach," *International Journal of Computer Engineering in Research Trends*, vol. 11, no. 1, pp. 1-8, 2024. [[CrossRef](#)] [[Publisher Link](#)]
- [6] J. Lampkins, Z. Huang, and Radwan, "Multimodal Perception for Dynamic Traffic Sign Understanding in Autonomous Driving," *Frontiers in Collaborative Research*, vol. 1, no. 1, pp. 22-34, 2023. [[Publisher Link](#)]
- [7] Daniel Parra-Ovalle, Carme Miralles-Guasch, and Oriol Marquet, "Pedestrian Street Behavior Mapping using Unmanned Aerial Vehicles. A Case Study in Santiago De Chile," *PLoS ONE*, vol. 18, no. 3, pp. 1-18, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [8] Taylor Li et al., "Pedestrian Behavior Study to Advance Pedestrian Safety in Smart Transportation Systems Using Innovative LiDAR Sensors," 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [9] Pannalal Boda, Y. Ramadevi, and M. Bhavsingh, "Leveraging Pre-Trained Vision for Enhanced Real Time Pedestrian Behavior Prediction at Zebra Crossings," *Frontiers in Collaborative Research*, vol. 1, no. 2, pp. 10-21, 2023. [[Publisher Link](#)]
- [10] Christian Brynning, A. Schirrer, and S. Jakubek, "Transfer Learning for Agile Pedestrian Dynamics Analysis: Enabling Real-Time Safety at Zebra Crossings," *Synthesis: A Multidisciplinary Research Journal*, vol. 1, no. 1, pp. 22-31, 2023. [[Publisher Link](#)]

- [11] Jun Yang et al., "Pedestrian Behavior Interpretation from Pose Estimation," *IEEE International Intelligent Transportation Systems Conference*, Indianapolis, IN, USA, pp. 3110-3115, 2021. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [12] Amir Rasouli, and Iuliia Kotseruba, "Pedformer: Pedestrian Behavior Prediction via Cross-Modal Attention Modulation and Gated Multitask Learning," *IEEE International Conference on Robotics and Automation*, London, United Kingdom, pp. 9844-9851, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [13] Chi Zhang, and Christian Berger, "Pedestrian Behavior Prediction Using Deep Learning Methods for Urban Scenarios: A Review," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 10, pp. 10279-10301, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [14] Shrutika Deokar, and Shridhar Khandekar, "Identification of Pedestrian Movement and Classification Using Deep Learning for Advanced Driver Assistance System," *International Conference on Augmented Intelligence and Sustainable Systems*, pp. 374-381, 2022. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [15] Ahmed Alhomoud et al., "Augmenting Real-Time Surveillance with EfficientDet a Leap Towards Scalable and Accurate Object Detection," *International Journal of Computer Engineering in Research Trends*, vol. 11, no. 2, pp. 9-17, 2024. [[CrossRef](#)] [[Publisher Link](#)]
- [16] Sheng-Chih Ho et al., "A Traffic Crash Warning Model for BOT E-Tolling Operations Based on Predictions Using a Data Association Framework," *Applied Science*, vol. 13, no. 10, pp. 1-13, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [17] Amirhossein Abdi, Seyedehsan Seyedabrishami, and Steve O'Hern, "A Two-Stage Sequential Framework for Traffic Accident Post-Impact Prediction Utilizing Real-Time Traffic, Weather, and Accident Data," *Journal of Advanced Transportation*, pp. 1-16, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [18] B. Pannalal, M. Bhavsingh, and Y. Ramadevi, "Enhancing Zebra Crossing Safety with Edge-Enabled Deep Learning for Pedestrian Dynamics Prediction," *International Journal of Computer Engineering in Research Trends*, vol. 10, no. 10, pp. 71-79, Oct. 2023. [[CrossRef](#)] [[Publisher Link](#)]
- [19] Liping Bao et al., "Learning Transferable Pedestrian Representation from Multimodal Information Supervision," *arXiv preprint arXiv:2304.05554*, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [20] Qi Zhang et al., "An Integrated Framework for Real-Time Intelligent Traffic Management of Smart Highways," *Journal of Transportation Engineering, Part A: Systems*, vol. 149, no. 7, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [21] Jia Huang, Alvika Gautam, and Srikanth Saripalli, "Learning Pedestrian Actions to Ensure Safe Autonomous Driving," *IEEE Intelligent Vehicles Symposium (IV)*, Anchorage, AK, USA, pp. 1-8, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [22] Yuxuan Wu et al., "Multi-Stream Representation Learning for Pedestrian Trajectory Prediction," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 3, pp. 2875-2882, 2023. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [23] Ling Zhao et al., "T-GCN: A Temporal Graph Convolutional Network for Traffic Prediction," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 9, pp. 3848-3858, 2019. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]
- [24] Hao Xue, Du Q. Huynh, Mark Reynolds, "Bi-Prediction: Pedestrian Trajectory Prediction Based on Bidirectional LSTM Classification," *International Conference on Digital Image Computing: Techniques and Applications*, pp. 1-8, 2017. [[CrossRef](#)] [[Google Scholar](#)] [[Publisher Link](#)]



ISSN:2147-6799

# International Journal of INTELLIGENT SYSTEMS AND APPLIED ENGINEERING

www.ijisae.org

## Inferring the Causal Relationships in Student Placements using Causal Machine Learning

<sup>1</sup>D. Naga Jyothi\*, Dr. Uma N. Dulhare<sup>2</sup>

Submitted:14/03/2024    Revised: 29/04/2024    Accepted: 06/05/2024

**Abstract:** This research proposes using causality models to analyse and infer student placement data between applications of Causal Machine Learning and Machine Learning for resolving different educational *does not equal Causation*. In traditional machine learning, the focus is often on predicting outcomes. However, causal machine learning goes beyond prediction by aiming to uncover cause-and-effect relationships. A review of causal inference in the presence of massive data sets is a rich and expanding field of contemporary research. The goal of causal inference is to understand how changes in one variable affect another, and to identify the underlying causal structure. The causal Inference which is the key concept for causal machine learning can be implemented using Directed Acyclic Graph (DAG). Through this paper we aim to provide some useful insights using 3 causal discovery tools (Data correlation, Discovery tool using Causal ML, Domain knowledge). We proposed a novel 3D framework (Data correlation, Discovery tool using Causal ML, Domain knowledge) and compared the merits of both manual and causal discovery tools. The causal graph obtained is checked for falsification using the DAG. The obtained graph needs to be informative and significance level ( $p$ -value  $< 0.05$ ) so that the DAG would be formed that represents relationships between the variables to understand and predict the effects of the system.

**Keywords:** Causal relationships, Causal discovery techniques, Directed Acyclic Graph (DAG), 3D Framework, Falsification, Causal Modelling.

### 1. Introduction

Despite all the hype surrounding AI, the majority of ML initiatives prioritise outcome prediction over causality analysis. Indeed, after several AI projects, It is realized that ML is great at finding correlations in data, but not causation. This problem severely restricts our ability to use Machine Learning for Decision Making.

Machine learning is a powerful tool to find patterns and to examine associations and correlations, particularly in large data sets [1]. Although the use of machine learning has led

It is reasonable to assume that the world model will be a critical component of AI systems in future. In traditional machine learning, the focus is often on predicting outcomes from data. However, causal machine learning goes beyond prediction by aiming to uncover cause-and-effect relationships between variables.

The commonly held belief that "correlation does not equal causation" refers to the fact that just because two variables are correlated, it does not mean that one causes the other. It is important to



## DEEP LEARNING BASED MALWARE DETECTION

T. SUSHMA<sup>1</sup>, SIRISHA NARKEDAMILI<sup>2</sup>, MADHAVA RAO CHUNDURU<sup>3</sup>, VADDEMPUDI SUJATHA LAKSHMI<sup>4</sup>, G. BALU NARASIMHA RAO<sup>5</sup>, PRABHAKAR KANDUKURI<sup>6</sup>

<sup>1</sup>Department of ECE, Prasad V Potluri Siddhartha Institute of Technology, Vijayawada, India

<sup>2</sup>Department of EEE, Aditya College of Engineering and Technology, Surampalem, India

<sup>3</sup>Department of CSA, Koneru Lakshmaiah Education Foundation, Vaddeswaram, India

<sup>4</sup>Department of Computer Applications, RVR&JC College of Engineering, Guntur, India

<sup>5</sup>Department of CSE, Vignan's Foundation of Science, Technology & Research, Guntur, India

<sup>6</sup>Department of AI&ML, Chaitanya Bharathi Institute of Technology, Hyderabad, India

tsushmaece@gmail.com , sirisha.narkedamilli@acet.ac.in, cmadhavarao@kluniversity.in,

sujathavdmpudi@gmail.com, balunarasimharao@gmail.com , prabhakarcs@gmail.com

### ABSTRACT

Malware detection is a critical aspect of cybersecurity, aiming to identify and mitigate malicious software designed to harm or exploit any programmable device or network. Traditional methods of malware detection, such as signature-based techniques, have limitations in dealing with the sophisticated and rapidly evolving nature of modern malware. This paper explores the application of deep learning, a subset of artificial intelligence, in enhancing malware detection capabilities. By leveraging deep learning models, which can automatically learn and extract features from data, we can improve detection accuracy and adapt to new, unseen malware. This research reviews various deep learning architectures and methodologies employed in malware detection, evaluates their effectiveness, and discusses future directions and challenges in the field.

**Keywords:** *Malware detection, deep learning, cybersecurity.*

### 1. INTRODUCTION

In the digital era, a significant number of computing devices have been impacted by malware. Malware, often known as malevolent software, is specifically designed to fulfill the harmful objectives of a malicious attacker. Malicious software, or malware, has the ability to infiltrate networks, cause harm to critical infrastructure, compromise the security of computers and smart devices, and unlawfully obtain confidential information [1].

The concept of an information society has developed due to the emergence of the Internet of Things (IoT) and its various uses. Nevertheless, the benefits of this industrial progress are impeded by security concerns, as hackers selectively target individual computers and networks to illicitly obtain confidential data for monetary purposes and disrupt operations [2]. These attackers employ harmful software, also

known as "malware," to exploit system weaknesses and present significant risks. Malware, also referred to as malicious software, is a type of computer software specifically designed to cause harm to an operating system [3]. The frequency of malware attacks has greatly risen due to the substantial changes in our daily contacts caused by the advancements in mobile technologies. Mobile devices connected to the Internet provide many services such as online learning, social networking, online banking, online shopping, and web browsing. Mobile devices have thus played a pivotal role and have transformed into an essential component of everyday life [4]. As of 2020, the global mobile device user count stands at 4.78 billion [5]. While mobile devices offer convenience to consumers, they are susceptible to virus infiltration and attacks due to their connection to online social networks and services. Mobile malware has the ability to masquerade as regular code and

thereafter modify any intended program in order to corrupt and impede the functioning of the system [5,6,7].

Google Play has implemented a permission-based approach as a security safeguard to prevent applications from accessing private data. This permission prompts users to install the program after considering the assets that have been accessed. Prior to proceeding with the installation, it is imperative that the users explicitly acknowledge and agree to the terms of the agreement. Regrettably, the Google Play technique does not provide complete protection for the customer since they often have a propensity to approve the agreement without thoroughly perusing the permission [5,8]. Another potential threat arises from the exploitation of profitable Android applications, as seen by the significant rise in the detection of Android malware, which increased more than tenfold from 2012 to 2018 [9]. In addition, a total of more than 12,000 new instances of Android malware were discovered per day during the year 2018. The newly discovered Android malware samples exhibit greater sophistication compared to those that emerged a few years ago, particularly in their ability to evade antivirus detection through coding and encryption. Additionally, there has been a significant increase in the spread of malware [10,11].

The utilization of machine learning in malware detection studies is becoming increasingly popular due to its effectiveness in achieving a high level of accuracy in detecting malware [12]. Prior research has employed machine learning (ML) algorithms, which have the ability to make decisions based on learned patterns from data. Machine learning refers to the idea of reducing the need for human involvement in computer systems [13]. Machine learning utilizes computer learning algorithms and historical data to make predictions. Supervised and unsupervised learning approaches [14,15] are utilized to examine the characteristics and monitor the model. In both scenarios, the machine acquires the ability to differentiate between harmful and harmless actions. In supervised learning, the machine learning model is provided with both the input data and the desired outputs. It then learns to accurately classify malware patterns as "malware" and normal behaviors as "normal". The training phase is iterated until the model achieves perfect accuracy in predicting all samples [5]. Various machine learning algorithms, such as support

vector machines (SVM) [16,17,18], K-nearest neighbor (KNN) [19,20], Bayesian estimates [21,22], genetic algorithms [23], have been employed to construct malware detection systems. Unsupervised learning approaches involve providing inputs without any predetermined targets, allowing the machine learning system to learn how to differentiate between malware and benign samples. Nevertheless, certain investigations integrated the approaches of supervised and unsupervised learning [24].

Malware detection is a crucial aspect of security that is closely linked to the legal, reputational, and economic interests of companies. Utilizing deep learning as a technique for developing and enhancing detection methods is an effective approach to address many challenges associated with malware detection. However, in the realm of deep learning, there are numerous complex factors that must be taken into account when considering detection strategies. Correlation-based feature selection, the dense layer model, and the LSTM model are three complex and symmetrical approaches that can significantly impact performance.

Two distinct datasets will be utilized in the ongoing research. One of the datasets contains a substantial quantity of records, whereas the other dataset comprises a significant number of predictors (attributes). The process of selecting the most optimal qualities will be employed in various situations to determine the most effective combination of features. The correlation between the target property "classification" will be utilized as the way for selecting features. The training phase will involve the utilization and comparison of Dense and LSTM models. Multiple training scenarios will be set up based on various feature selection criteria, splitting criteria, and dataset topologies. Our primary innovation lies in using the efficacy of deep learning and feature selection techniques in the domain of malware detection to construct a resilient, high-performing, computationally efficient malware detection system.

## 2. METHODOLOGY

This research employs both static and dynamic analysis methods using deep learning models. The dataset comprises a mix of known malware samples and benign software, sourced from public repositories like VirusTotal and Maling dataset. Multiple deep learning (DL) methods are suggested and employed in this

work. To train the deep learning models using the two chosen datasets, it is necessary to preprocess these datasets. This preprocessing phase involves encoding (numbering) the classification (target) columns and handling any special characters or missing values. Due to the distinct nature of the two datasets, the preparation stages will vary. Once the datasets have undergone preprocessing, they are divided into separate training and test sets. Feature selection is conducted prior to the training process in certain training scenarios to reduce data dimensionality and computational time.

Subsequently, the DL models will be constructed and trained using several training scenarios, encompassing diverse splitting criteria, distinct DL architectures, and the option of feature selection. Figure 1 depicts the recommended technique for both datasets.

The objective of feature selection is to identify the most optimal characteristics relevant to the topic being examined, with the purpose of minimizing computational time. Our study proposes a correlation-based technique to address the issues of large dimensionality and long processing time. This approach also aims to pick the most effective combinations of features, hence improving the performance of the training and evaluation process.

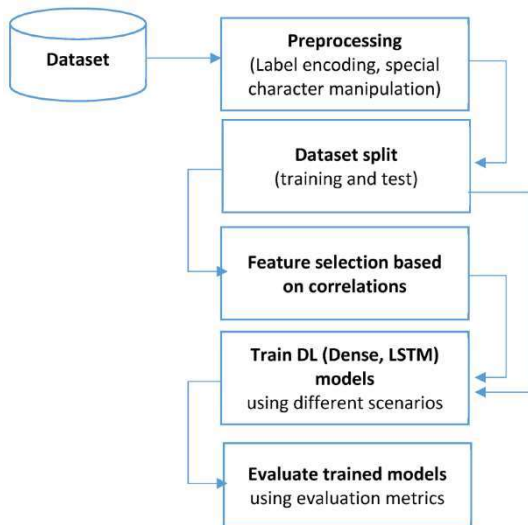


Figure 1. Illustrates the methodology that is being suggested

Next, a compilation of probable columns to be dropped is generated. Various selection scenarios can be generated as the correlation spans from 0 to 1. The selection process is determined by the desired number of columns. We will extract the K necessary features and

discard the remaining ones. For the second dataset, the identical method will be utilized, with the exception of the selection phase. Columns will be eliminated in the second dataset based on specific correlation thresholds (T), given that there are 214 columns in total. In the second dataset, the number of selected features is contingent upon the chosen threshold, unlike the first dataset where the threshold is not specified.

### 3. RESULTS AND DISCUSSION

This section presents and discusses the findings of the experiments done to assess the effectiveness of the autoencoder-based malware detection approach. The current section is dedicated to providing pertinent information on the experimental setup. It is divided into three sections to address the following aspects: the configuration of the experiment setup, the gathering of data, and the specifics of the training process. To obtain additional details regarding the experimental setting, go to table 1. The tests were conducted on a machine equipped with an Intel Core™ i5-8300 processor, 16GB of RAM, and a GeForce GTX 1060 MQ graphics card. The computer operated on a 64-bit iteration of the Windows 10 operating system. In our programming, we employed Keras, Tensorflow 2.1, and Python 3.7. We categorized the datasets according to their intended purpose. (1) Dataset-1 consists of 8,121 malicious programs and 2,000 benign programs. It is utilized for training and evaluating AE-1 models. The AE-2 model is trained, validated, and tested using Dataset-2, which consists of 8121 dangerous applications and 7015 safe ones. The AE-2 model is tested using Dataset-3, which consists of 5,384 malicious applications and 5,000 safe programs, to evaluate its ability to detect unfamiliar malware. It is important to note that when we divided Dataset-2 and Dataset-3, we intentionally incorporated older software samples in Dataset-2 for the purpose of training, such as malware from 2016.

In Dataset-3, we included more recent releases, such as those from 2017 and 2018. Simulating the condition of identifying newly published software samples will facilitate future analysis of the model's performance. In order to evaluate the autoencoder's ability to reconstruct feature images, we utilize the AE-1 network. The specific attributes of the AE-1 model are presented in Table 2. During the training process, we utilize the Adam optimization technique with a total of 100 epochs and a learning rate of 1e-4.

The AE-1 network undergoes training using the DTrain dataset and subsequently undergoes testing using the DTest\_mal and DTest\_benign datasets, which contain malicious and benign software, respectively. In order for a test set to have a low reconstruction error, the new input must be similar to the input of the training dataset. Conversely, if the new inputs deviate from the inputs used in the dataset during training, a noticeable reconstruction error will be observed in this test set. The primary objective of our experiment is to investigate the significant disparity in error data produced by these two test sets after AE-1. In practical terms, the significant duplication features in the software dataset and the distinct functional traits exhibited by malware families in the malware dataset can result in experimental outcomes showing substantial fluctuations. This is because our hypothesis is founded on the notion that malware is universally similar, while benign software is not. Consequently, we place less importance on the exact errors exhibited by the two test sets and instead focus more on the comparative disparities between them. The responsibility of evaluating the performance of the detection model lies with the AE-2 network. We partitioned Dataset-2 into two equal parts, allocating 80% for training and 20% for testing. During the training process, we employed k-fold cross-validation with a value of k equal to 6 in order to train and validate the training set. Consequently, we allocated 5/6 of the training set for training purposes and reserved 1/6 for validation. We conducted this procedure on six occasions prior to calculating the average. The test set is used for testing purposes during the entire testing procedure. The duration of training is quantified in units of minutes. The training of AE-2 utilized the Adam optimization technique with a learning rate of 0.0001 and 100 epochs. We evaluate the effectiveness of this strategy by analyzing the overall error distribution in both malicious and benign reconstructions of malware images. Figure 2 shows the error distributions of the combined test sets. The Y-axis indicates the normalized reconstructed error value generated by each program following the encoder network. The error value of each pixel point corresponding to the malware feature image is aggregated and subsequently divided by the total to achieve image normalization. The line statistics graph illustrates the general error trend of DTest\_mal with a blue line, whereas the overall error trend of DTest\_benign is represented by a yellow line. The inherent unpredictability of the dataset plus

the redundancy of the software files result in a non-zero error. Figure 5 illustrates a significant disparity in the average error values between the two datasets. The blue line indicates a consistent and steady error trend for the malware dataset, while the yellow line represents an erratic and fluctuating error trend for the benign software test set. This supports our perspective.

Based on this experiment, we can show that the automatic encoder can identify complex characteristics of both harmless and harmful software and successfully reconstruct the pre-processed malware data. Next, we proceed to carry out the task of differentiating between harmful and benign software.

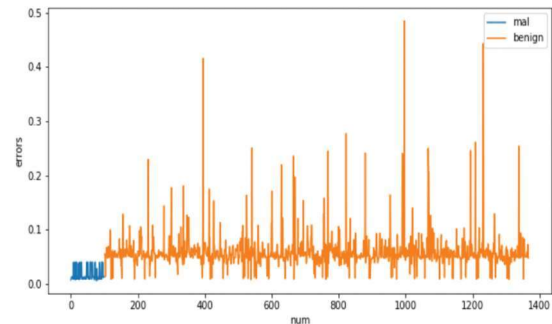


Figure 2. Mistake in reconstruction for two sets of data.

The evaluation of the autoencoder model is conducted using many measures, such as accuracy, precision, recall, F1-score, false positive rate, and false negative rate. These metrics offer a holistic perspective on the model's capacity to accurately detect instances of malware while minimizing both false positives and false negatives. The results are compared with conventional signature-based methods to emphasize the possible enhancements provided by the autoencoder methodology. Signature-based approaches are intrinsically constrained by their dependence on pre-established patterns, rendering them vulnerable to evasion by polymorphic and metamorphic malware. The autoencoder's capacity to acquire knowledge from the inherent characteristics of data without pre-established patterns situates it as a more flexible and adaptable solution. The ROC curves depicted in Figure 3 illustrate the impact of the model on the training set. It is evident that the model demonstrates a consistently reliable performance on the training set. The ROC curves depicted in Figure 4 illustrate the model's performance on the test set, specifically Dataset-2. It is evident that our model surpasses the other

two in performance. In order to thoroughly evaluate the ability of our model to detect previously undiscovered malware, we utilize Datasets-3 as the test set for AE-2. The ROC curves are displayed in Figure 5, revealing that our model exhibits commendable accuracy and a certain degree of viability in detecting previously undetected malware. However, it also exhibits certain limitations as the software evolves over time.

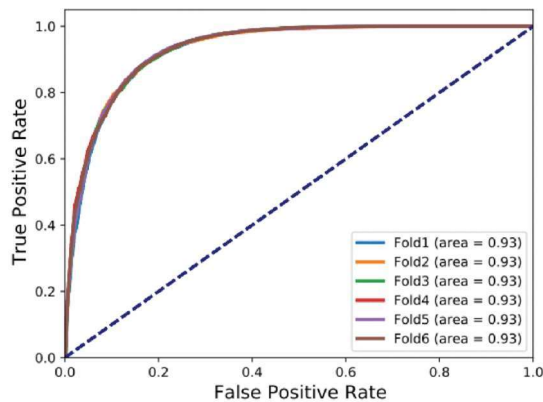


Figure 3. The ROC curve of AE-2 on training set.

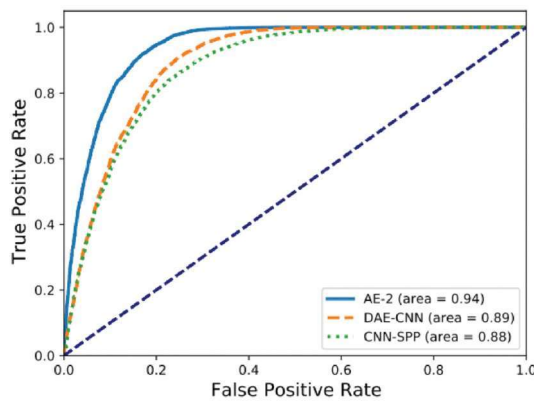


Figure 4. The ROC curve of different models on the test set.

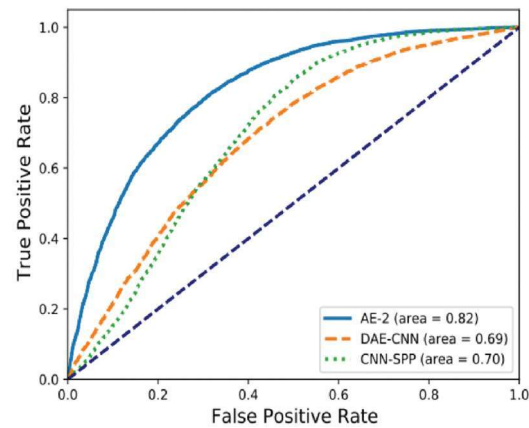


Figure 5. The ROC curve of different models on the unseen software

The results and discussion section illuminates the performance of the autoencoder-based malware detection approaches, offering valuable insights into its strengths, limitations, and implications for the broader cybersecurity landscape.

#### 4. CONCLUSION

The findings of this study emphasize the capacity of autoencoders to enhance the skills of malware detection. Autoencoders provide a promising approach to improving the adaptability and effectiveness of cybersecurity defenses against emerging malware threats by adopting the dynamic and unsupervised learning paradigm. The study findings obtained from the evolving digital landscape contribute to the ongoing effort to develop new and robust solutions for protecting digital ecosystems. The testing results confirm the effectiveness of our proposed approach, which entails converting the bytecode of each software method into a grayscale image that graphically depicts the characteristics of a software sample. Our approach exhibits a notably higher level of accuracy in identifying malicious software compared to methods built using traditional machine learning algorithms. Our technique exhibits decreased training and detection durations when compared to competing malware detection systems that depend on deep learning models. The text outlines suggestions for future research approaches, including investigating ensemble methods that combine autoencoders with other deep learning architectures, including temporal factors to improve dynamic malware detection, and utilizing adversarial training to boost the robustness of models. These recommendations

are intended to provide guidance for future investigations in the continual pursuit of developing more efficient and adaptable malware detection systems.

## REFERENCES

- [1]. Rathore, H.; Agarwal, S.; Sahay, S.; Sewak, M. Malware detection using machine learning and deep learning. In Proceedings of the International Conference on Big Data Analytics, Seattle, WA, USA, 10–13 December 2018; pp. 402–411. [Google Scholar]
- [2]. Nasif, A.; Othman, Z.; Sani, N.S. The deep learning solutions on lossless compression methods for alleviating data load on IoT nodes in smart cities. *Sensors* 2021, 21, 4223. [Google Scholar] [CrossRef] [PubMed]
- [3]. Vinayakumar, R.; Alazab, M.; Soman, K.; Poornachandran, P.; Venkatraman, S. Robust intelligent malware detection using deep learning. *IEEE Access* 2019, 7, 46717–46738. [Google Scholar] [CrossRef]
- [4]. Singh, A.; Kumar, R. A two-phase load balancing algorithm for cloud environment. *Int. J. Softw. Sci. Comput. Intell.* 2021, 13, 38–55. [Google Scholar] [CrossRef]
- [5]. Mat, S.R.T.; Razak, M.A.; Kahar, M.; Arif, J.; Firdaus, A. A Bayesian probability model for Android malware detection. *ICT Express* 2022, 8, 424–431. [Google Scholar] [CrossRef]
- [6]. Yen, S.; Moh, M.; Moh, T.-S. Detecting compromised social network accounts using deep learning for behavior and text analyses. *Int. J. Cloud Appl. Comput.* 2021, 11, 97–109. [Google Scholar] [CrossRef]
- [7]. Shabudin, S.; Sani, N.; Ariffin, K.; Aliff, M. Feature selection for phishing website classification. *Int. J. Adv. Comput. Sci. Appl.* 2020, 11, 587–595. [Google Scholar] [CrossRef]
- [8]. Liu, C.-H.; Zhang, Z.-J.; Wang, S.-D. An android malware detection approach using Bayesian inference. In Proceedings of the 2016 IEEE International Conference on Computer and Information Technology (CIT), Nadi, Fiji, 8–10 December 2016; pp. 476–483. [Google Scholar]
- [9]. GDATA Mobile Malware Report—No let-up with Android malware. 2019. Available online: <https://www.gdatasoftware.com/news/2019/07/35228-mobile-malware-report-no-let-up-with-android-malware> (accessed on 22 November 2022).
- [10]. Qiu, J.; Zhang, J.; Luo, W.; Pan, L.; Nepal, S.; Xiang, Y. A survey of android malware detection with deep neural models. *ACM Comput. Surv.* 2020, 53, 1–36. [Google Scholar] [CrossRef]
- [11]. Sihwail, R.; Omar, K.; Ariffin, K.A.Z. An effective memory analysis for malware detection and classification. *Comput. Mater. Contin.* 2021, 67, 2301–2320. [Google Scholar] [CrossRef]
- [12]. Mat, S.R.T.; Razak, M.A.; Kahar, M.; Arif, J.; Mohamad, S.; Firdaus, A. Towards a systematic description of the field using bibliometric analysis: Malware evolution. *Scientometrics* 2021, 126, 2013–2055. [Google Scholar] [CrossRef]
- [13]. Bassel, A.; Abdulkareem, A.; Alyasseri, Z.; Sani, N.; Mohammed, H.J. Automatic Malignant and Benign Skin Cancer Classification Using a Hybrid Deep Learning Approach. *Diagnostics* 2022, 12, 2472. [Google Scholar] [CrossRef]
- [14]. Jerlin, M.A.; Marimuthu, K. A new malware detection system using machine learning techniques for API call sequences. *J. Appl. Secur. Res.* 2018, 13, 45–62. [Google Scholar] [CrossRef]
- [15]. Abdallah, A.; Ishak, M.K.; Sani, N.S.; Khan, I.; Albogamy, F.R.; Amano, H.; Mostafa, S.M. An Optimal Framework for SDN Based on Deep Neural Network. *Comput. Mater. Contin.* 2022, 73, 1125–1140. [Google Scholar] [CrossRef]
- [16]. Han, H.; Lim, S.; Suh, K.; Park, S.; Cho, S.; Park, M. Enhanced android malware detection: An svm-based machine learning approach. In Proceedings of the 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), Busan, Republic of Korea, 19–22 February 2020; pp. 75–81. [Google Scholar]
- [17]. Singh, P.; Borgohain, S.; Kumar, J. Performance Enhancement of SVM-based ML Malware Detection Model Using Data Preprocessing. In Proceedings of the 2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET), Patna, India, 24–25 June 2022; pp. 1–4. [Google Scholar]

- [18]. Droos, A.; Al-Mahadeen, A.; Al-Harasis, T.; Al-Attar, R.; Ababneh, M. Android Malware Detection Using Machine Learning. In Proceedings of the 2022 13th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 21–23 June 2022; pp. 36–41. [Google Scholar]
- [19]. Baldini, G.; Geneiatakis, D. A performance evaluation on distance measures in KNN for mobile malware detection. In Proceedings of the 2019 6th international conference on control, decision and information technologies (CoDIT), Paris, France, 23–26 April 2019; pp. 193–198. [Google Scholar]
- [20]. Assegie, T.A. An optimized KNN model for signature-based malware detection. Tsehay Admassu Assegie. *Int. J. Comput. Eng. Res. Trends (IJCERT)* 2021, 8, 2349–7084. [Google Scholar]
- [21]. Castillo-Zúñiga, I.; Luna-Rosas, F.; Rodríguez-Martínez, L.; Muñoz-Arteaga, J.; López-Veyna, J.; Rodríguez-Díaz, M.A. Internet data analysis methodology for cyberterrorism vocabulary detection, combining techniques of big data analytics, NLP and semantic web. *Int. J. Semant. Web Inf. Syst.* 2020, 16, 69–86. [Google Scholar] [CrossRef]
- [22]. Yilmaz, A.B.; Taspinar, Y.; Koklu, M. Classification of Malicious Android Applications Using Naive Bayes and Support Vector Machine Algorithms. *Int. J. Intell. Syst. Appl. Eng.* 2022, 10, 269–274. [Google Scholar]
- [23]. Yildiz, O.; Doğru, I.A. Permission-based android malware detection system using feature selection with genetic algorithm. *Int. J. Softw. Eng. Knowl. Eng.* 2019, 29, 245–262. [Google Scholar] [CrossRef]
- [24]. Arora, A.; Peddoju, S.; Chouhan, V.; Chaudhary, A. Hybrid Android malware detection by combining supervised and unsupervised learning. In Proceedings of the 24th Annual International Conference on Mobile Computing and Networking, New Delhi, India, 29 October–2 November 2018; pp. 798–800. [Google Scholar]

# Event-based Smart Contracts for Automated Claims Processing and Payouts in Smart Insurance

Dr Araddhana Arvind Deshmukh<sup>1</sup>, Prabhakar Kandukuri<sup>2</sup>, Dr Janga Vijaykumar<sup>3</sup>,  
Anna Shalini<sup>4</sup>, Dr. S. Farhad<sup>5</sup>, Elangovan Muniyandy<sup>6</sup>, Dr. Yousef A. Baker El-Ebiary<sup>7</sup>

Professor, School of Computer Science & Information Technology (Cyber Security),  
Symbiosis Skill and Professional University, Kiwale, Pune, India<sup>1</sup>

Professor, Department of Artificial Intelligence and Machine Learning,  
Chaitanya Bharathi Institute of Technology - Hyderabad, India<sup>2</sup>

Associate Professor, Dept of CSE (AI&ML), Balaji Institute of Technology and Science, Narsampet, India<sup>3</sup>

Research Scholar, Dept of English, Koneru Lakshmaiah Education Foundation,  
Green Fields, Vaddeswaram, Guntur, Andhra Pradesh, India<sup>4</sup>

Associate Professor, Dept. of English, Koneru Lakshmaiah Education Foundation,  
Vaddeswaram, Guntur, Andhra Pradesh, India<sup>5</sup>

Department of R&D, Bond Marine Consultancy, London EC1V 2NX, UK, Department of Biosciences, Saveetha School of  
Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, India<sup>6</sup>

Faculty of Informatics and Computing, UniSZA University, Malaysia<sup>7</sup>

**Abstract**—The combination of blockchain technology and smart contracts has become a viable way to expedite claims processing and payouts in the quickly changing insurance industry. Enhancing efficiency, transparency, and reliability for the industry may be achieved by automating certain procedures and initiating them on predetermined triggers, smart contracts that is event-based. Conventional insurance procedures can be laborious, slow, and prone to human mistake, which can cause inefficiencies and delays in the resolution of claims. This research proposes a simplified system that automates the whole claims process from submission to reimbursement by utilizing blockchain technology and smart contracts. The suggested method does away with the requirement for human claim filing by having policyholders' claims automatically triggered by predetermined occurrences. These occurrences might be anything from medical emergencies to natural calamities, enabling prompt and precise claim start. The whole claims process is managed by smart contracts that are programmed with precise triggers and conditions, guaranteeing transaction immutability, security, and transparency. Moreover, reimbursements are carried out automatically after the triggering event has been verified, disregarding conventional bureaucratic processes and drastically cutting down on processing times. This strategy decreases the possibility of fraud and disagreement while also improving operational efficiency by combining self-executing contracts with decentralized ledger technology. Insurance companies and policyholders will both eventually profit from an accelerated, transparent, and reliable claims processing procedure thanks to the use of event-based smart contracts. A Python-implemented system achieving 97.6% accuracy using the proposed method, demonstrates its efficacy and reliability for the given task.

**Keywords**—Blockchain technology; smart contracts; event-based triggers; automated claims processing; transparency and trustworthiness

## I. INTRODUCTION

Private insurance businesses act as a central organization to give advantages to policyholders. They offer worth by using historic information and mathematical procedures to determine whether premiums are going to be adequate for covering predicted claims. Furthermore, authorities can control these companies in order to ensure enough funding. Insurance firms are currently losing a significant amount of money as a result of claims leakage. Illegal claims are a massive and expensive issue for insurance firms, possibly resulting in trillions of dollars in unwarranted spending every year [2]. Traditional policy approaches for detecting fraud are complex and time-consuming. They mostly rely on expert inspection, adjusters as well, and specialized investigative services. Manual inspection faces extra costs and yields erroneous findings. Furthermore, delayed decisions may result in additional losses for the insurance firms. The insurance business administrators auto-insurance processing of claims using information gathered from several domains, including police, county administration, insurance representatives, and medical professionals [3]. These businesses work together to communicate multi-source data, which is crucial enabling insurance firms to correctly assess customer claims. Yet, the majority of existing claim processing procedures are laborious and time-consuming because to a lack of automated methods to perform information collection/analysis, along with technology to make reliable decisions [4]. To enhance the effectiveness and adaptability of insurance claim the process, it is necessary to integrate automated processes and trust administration mechanisms at an application level [5]. The excessive number of false claims given out by motor insurance firms has resulted in price hikes of several hundred dollars to counteract the false payouts, reducing insurance company profitability as well as the level of operations [6]. As a result, there exists an urgent need to provide quick and effective solutions for identifying fraud, risk assessment, and



safe storage of information that strike a perfect equilibrium among customer private information safeguarding, loss prevention savings, and expenditure on false alarm identification (Cousaert, Vadgama, and Xu 2022). Recommend creating a successful framework for insurance companies to address such difficulties [7].

Intelligent contracts may streamline numerous common operations in the P&C insurance industry, including policy issuance and claim management. For example, parameterized insurance policies can initiate payments whenever predetermined conditions are satisfied, such as in the case of a natural disaster [8]. It may also assess all payment choices to determine which one is optimal. This technology eliminates the requirement for middlemen and increases effectiveness, giving policyholders access actual time replies which are not impeded by delays in insurance claims [9]. This study examines the notion of automation claims handling and payouts driven by event-driven intelligent contracts in the insurance business. The technological infrastructure necessary to construct a system like this, includes the selection of blockchain platform, programming languages, and architectural considerations for smart contract implementation [10]. Research demonstrate the flow of data and actions across the system, emphasizing the seamless integration of event-driven triggers for automating the claims handling workflow [11]. Its technological infrastructure necessary to construct any of these systems, in addition to the selection regarding the blockchain platform, programming languages, and architectural considerations for smart contracts .Demonstrate the flow of data and operations across the system, emphasizing the effortless incorporation of based on events triggering for automating the claims handling process [12].

The Current solutions based on blockchain employ intelligent agreements to enhance the transfer of assets, restrict fraud, and decrease administrative expenses. However, they do not address collaborative insurance, allowing individuals who have comparable characteristics to safeguard each other in an increased favorable, reasonable, and open way [13]. To illustrate whether event-based intelligent agreements might transform the claims handling procedure by thoroughly examining its technological foundations and operational ramifications, benefiting insurance companies, policyholders, and various other stakeholders equally. With adopting this novel strategy, insurance firms may achieve unprecedented levels of efficiency, openness, and satisfaction with customers, bringing in an entirely novel phase of insurance claim administration [14].

Key contributions are as follows:

- By automating claims processing through event-based smart contracts, the system eliminates manual submission processes, reducing administrative burdens and streamlining operations.
- Blockchain technology ensures transparency and immutability of transactions, providing a clear audit trail for all stakeholders involved.
- Predefined events trigger claims automatically, enabling quick initiation upon the occurrence of

insured events such as natural disasters or medical emergencies. This swift response enhances customer satisfaction and reduces delays in claim settlements.

- The inherent security features of blockchain technology, combined with self-executing smart contracts [1] minimize the potential for fraudulent activities in the claims process.
- Automated payouts upon verification of triggering events bypass traditional bureaucratic procedures, significantly reducing processing times.
- By streamlining processes and reducing manual intervention, insurers can realize cost savings and operational efficiencies.

The remaining section of this work is structured as follows: Section II covers similar work and a full evaluation of it. Section III offers details on the problem statement. Section IV provides a detailed discussion of the suggested method. Section V presents and examines the results of the tests, as well as a comprehensive comparison of the proposed technique to current standard procedures. Section VI, the last section, represents where the paper is finished.

## II. RELATED WORKS

The existing health insurance claims procedure has issues with inefficiency and complexity. Whenever a patient files a health insurance claim, he or she must first visit the medical facility to obtain a diagnostic certification and being received, and finally submit the required application documentation to the insurer. The person will not get compensation until the company completes its verification procedure via the patient's clinic. Research can use the technology of blockchain to better the existing situation. Blockchain innovation may successfully open up avenues for communication between insurance companies and healthcare providers, increase industrial integrating, and improve healthcare firms' capacity to access data. This study uses blockchain and smart contract technology to boost the progress of Internet healthcare. First, blockchain and smart contracts technology may effectively handle the problem of web-based verification. In addition, it contributes to better monitoring. Finally, it helps to solve risk management issues. Finally, it promotes efficient anti-money laundering. The suggested approach meets a number of safety criteria: mutual verification of identities and the non-rep among all of both roles, along with additional significant the blockchain relies safety concerns. In the case of a conflict provide an arbitration system to distribute duties. The effective deployment of the blockchain system in the insurance sector necessitate the development of strong publicly accessible infrastructure (PKI), partnerships between healthcare providers for offering electronic health records (EMR), as well as money alliances for expressing consumer financial data, that could create practical and legal obstacles in some countries [15].

The insurance sector, firms have implemented substantial and fundamental modifications to update their basic processes, making operations simpler and quicker for customers and enterprises. To service more clients while enhancing the total

client experience across all contact points, organizations are seeking to shift out of standalone transactional systems and towards contextually engagement systems. Several insurance companies currently use some form of automation, including scanning, uploading papers for the process, or automating bank transfer activities. However, occasionally this might result in inadequate results or delayed procedures. Robotic Process Automation (RPA) is the employing of computer programs robots to execute business operations that would normally be performed by humans. RPA can help companies accomplish their business goals while utilizing existing technology and increasing the returns on prior and ongoing transformational expenditures. Insurers may utilize RPA to analyze large amounts of complicated data at greater rates and in less time. RPA is poised to assist claiming businesses develop and improve their results in the age of technology by increasing automated processes, efficiency, and concentration for claim experts. Companies with superior outsourced capabilities have widened their concentration on automating to save labor expenses and streamline procedures. The following has generated an emerging RPA industry that is expected to expand significantly. RPA's drawbacks includes being unable to perform activities that need complicated making choices or mental skills, as well as its dependence on organized information and repeated procedures, that might not apply to all circumstances or sectors [16].

Dhieb et al. [17] propose safe and automatic healthcare system architecture that eliminates human intervention, protects insurance operations, notifies and educates concerning dangerous consumers, identifies forged claims, and decreases the financial loss for the insurance industry. Subsequently introducing the blockchain relies system for enabling secure transactions as well as information offering between various agents who communicate inside the insurance company network, that research suggest employing the xtreme gradient boosting (XGBoost) artificial intelligence method for the formerly mentioned insurance companies and comparing its efficacy to that of other cutting-edge algorithms. The findings show that when implemented to an automobile insurance dataset, Boost outperforms other present-day learning methods. When it comes to identifying false claims, it outperforms the decision tree algorithms by 7% on average. The findings show that whenever deployed to an automobile insurance dataset, XGboost outperforms alternative present-day learning methods. Whenever it comes to identifying false claims, it outperforms the decision tree models by 7% on average. In addition, present an online educational approach to autonomously cope with real-time modifications to the insurance network, as well as demonstrate that it beats other online cutting-edge method. At last use the hyper ledger networks fabric composer and the built neural network modules to construct and replicate the machine learning algorithms and bit coin architecture. Throughout the coming years, company are going to concentrate on improving the proposed framework and introducing artificial intelligence (AI) products targeted to various insurance services.

The insurance sector relies largely on a number of activities carried out by different organizations, including insurers, insured's, and third-party service providers. The

growing competitive climate is driving insurance businesses to adopt innovative technology to solve a variety of issues, including an absence of confidence, openness, and economic uncertainty. For this purpose, blockchain is being employed as a new technology for accessible and safe information preservation and transfer. Loukil et al. [18] propose CioSy, an integrated a blockchain-based healthcare platform that monitors and processes insurance activities. To the greatest extent of understanding, current processes do not take cooperative insurance into account while aiming for a computerized, clear, and tamper-proof solution. CioSy intends to use smart contracts to automate the processing of insurance policies, claims, and payments. For validation reasons, an experimental prototype is created on the Ethereum blockchain. The findings from experiments suggest that the suggested strategy is viable and cost-effective. In the future, research hopes to give a formal privacy demonstration for the suggested paradigm. In addition, intend to investigate the feasibility of deploying the funds gathered by an insurance pooling utilizing blockchain-based technology with the goal to encourage bankers and insurance organizations to join a proposed collaboration healthcare system.

Traditional claims handling procedures are inadequate for the current world, which has an expanding fleet of cars and an equal amount of incidents. Fernando et al. [19] suggest a fresh proposal for automating the financial services industry's laborious operations. Its provided approach is made up of three primary elements: re-identifying the car's model and year, identifying the harmed automotive part, kind, and extent, and computing a precise repair cost utilizing damages part recognition. Simplify the recording process by detecting important fields from the user's voice input. This guarantees that all parties participating in the procedure benefit from the proposed system. The presented solutions were developed utilizing Artificial Intelligence approaches, namely CNN models and natural language processing techniques. The initiative's planned developments for the future include improving the ASR to detect more fields linked to completing out the initial claim seeking form as well as including additional regional dialects. The given technique is capable of recognizing one type of harm in a picture. This may be enhanced to identify multiple kinds of harm in a picture as technology for computer vision evolves. These improvements will improve the overall efficiency of the system in the years to come.

Machine learning or data mining algorithms may be utilized for forecasting future management and are thus considered strong tools. Data mining has recently become increasingly significant for obtaining essential data in the healthcare industry. Health insurance costs are critical in the development of healthcare institutions. In order to offer improved healthcare services, it is critical to anticipate the cost of medical insurance that constitutes one of the opportunities for improving healthcare facilities. Dutta et al. [20] addresses projecting the cost of medical coverage, which must be provided by the individual receiving medical care. To accomplish the best predictions examination, several data mining regression techniques are used, including decision trees, random forests, polynomial regression, and regression

using linear models. A contrast was made among the actual and expected expenditures for the predictions premiums, and a graph was created on this foundation to help us identify the optimum method of regression for insurance policy prediction. One constraint is the possible complexity and technical needs of adopting sophisticated neural network algorithms such as Bi-LSTM, that might necessitate extensive knowledge and computing power. Another drawback is the absence of insurance-related information, which restricts the research to a small dataset and could restrict the ability to generalize of the findings. Every method is evaluated to determine the most appropriate solution.

While several studies have highlighted the potential of emerging technologies such as blockchain, robotic process automation (RPA), machine learning, and data mining in revolutionizing the health insurance claims process, there remain significant limitations across these works. Firstly, while blockchain offers secure data sharing, its implementation may face challenges related to infrastructure development and legal obstacles. Additionally, the reliance on structured data and repetitive processes in RPA may limit its applicability in complex decision-making scenarios. Moreover, the effectiveness of machine learning algorithms like XGBoost and data mining techniques in predicting insurance costs is constrained by the availability of comprehensive datasets and computational resources. Furthermore, the complexity of advanced neural network algorithms may hinder their adoption, while the lack of insurance-specific information can restrict the generalizability of findings. These limitations underscore the need for further research to address technical, data-related, and practical challenges in leveraging emerging technologies for enhancing the efficiency and effectiveness of health insurance processes.

The existing health insurance claims procedure is plagued by inefficiency and complexity, requiring patients to visit medical facilities to obtain diagnostic certification and then submit documentation to insurers, resulting in delayed compensation. To address these issues, the current research proposes utilizing blockchain and smart contract technology. This technology facilitates communication between insurance companies and healthcare providers, enhances industrial integration, and improves healthcare firms' access to data. The proposed approach aims to boost the progress of Internet healthcare by effectively handling web-based verification, improving monitoring, and addressing risk management issues. Dhieb et al. proposed a safe and automatic healthcare system architecture that eliminates human intervention, protects insurance operations, identifies forged claims, and decreases financial loss. They suggest employing the XGBoost artificial intelligence method for insurance companies, which outperforms other algorithms in identifying false claims. Similarly, Loukil et al. proposed CioSy, a blockchain-based healthcare platform that automates insurance policies, claims, and payments through smart contracts. Fernando et al. suggest automating financial services operations, including car damage assessment for insurance claims, using Artificial Intelligence approaches. Dutta et al. on predicting medical insurance costs, utilizing various data mining regression techniques. These earlier

studies provide a comprehensive framework for the current research on Event-Based Smart Contracts for Automated Claims Processing and Payouts in Smart Insurance. The proposed system will leverage blockchain technology and smart contracts to automate and streamline the insurance claim process, enhancing efficiency, transparency, and security. By integrating findings from previous research, the proposed system will significantly contribute to solving the inefficiencies and complexities of the existing health insurance claims procedure.

### III. PROBLEM STATEMENT

The current insurance claims procedure is plagued by inefficiencies and complexities, requiring physically visit facilities for certification before submitting paperwork to insurance firms. This cumbersome process leads to delays in compensation and poses challenges in data verification and risk management [18]. However, emerging technologies like blockchain and Robotic Process Automation (RPA) offer promising solutions to streamline these operations. Blockchain can facilitate secure communication between insurance companies and healthcare providers, while RPA can automate repetitive tasks, improving efficiency and accuracy. Additionally, the integration of artificial intelligence (AI) algorithms, such as XGBoost, enhances fraud detection and claim processing speed. Despite these advancements, there remain challenges in implementing these technologies, including the need for robust infrastructure and data privacy considerations [16]. Hence, there is a pressing need for innovative solutions like CioSy, a blockchain-based healthcare platform, which automates insurance processes through smart contracts, ensuring transparency and reliability. Furthermore, leveraging AI techniques like convolutional neural networks (CNN) and natural language processing (NLP) can further enhance claims handling by automating tasks like damage assessment and form completion. Ultimately, adopting these technologies can revolutionize the insurance sector, making processes more efficient, transparent, and customer-centric. The Novel method Automated Claims Processing and Payouts Triggered by Event-Based Smart Contracts is proposed.

### IV. PROPOSED METHOD AUTOMATED CLAIMS PROCESSING AND PAYOUTS TRIGGERED BY EVENT-BASED SMART CONTRACTS

The suggested event-driven architecture insurance claim procedure follows a certain set of phases in its approach. First, pertinent data on insurance plans, applicants, and triggering events are gathered through data collection and preprocessing. After that, this data is examined and plotted to reveal trends and patterns that might guide the creation of smart contracts. Subsequently, blockchain technology is utilized for safe transactions and smart contract implementation in the automatic medical insurance claims processing system. Automate the claims procedure and ensure speed and transparency, smart contracts are configured with certain triggers and criteria. When certain events occur, such as natural catastrophes or medical emergencies, smart contracts automatically start the claims procedure. This entails confirming the legitimacy of the claim and streamlining the reimbursement procedure without requiring human

involvement. Utilizing blockchain technology, this technique places a strong emphasis on guaranteeing the confidentiality and integrity of the claims process. In addition, the system is

routinely optimized and monitored to preserve its efficacy and efficiency in managing insurance claims.

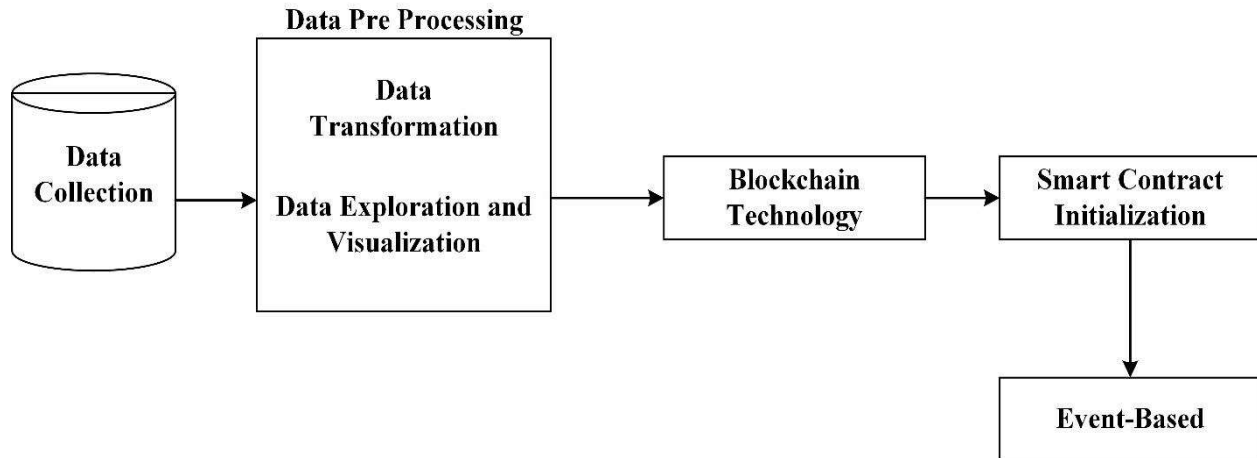


Fig. 1. Automated claims processing and payouts triggered by event-based smart contracts.

Fig. 1 displays a flowchart of a data-driven workflow using blockchain technology. The procedure includes five steps: data gathering, data pre-processing, blockchain technology, smart contract activation, and event-based process.

#### A. Data Collection

The "Health Insurance Dataset - EDA" available on Kaggle offers a comprehensive exploration of health insurance data, comprising information on 1338 US health insurance customers. The dataset includes features such as age, gender, body mass index (BMI), number of children, smoking status, region, and insurance charges. It serves to facilitate analysis on factors affecting insurance charges, prediction of new charges, and comparison of plans across different regions. Key inquiries encompass the impact of age, smoking status, and number of children on insurance charges, identification of regions with the highest or lowest average charges, and examination of BMI distribution across regions. This dataset is valuable for understanding insurance pricing and trends [21].

#### B. Data Pre Processing

Data preprocessing involves cleaning and transforming raw data to enhance its quality and usability for analysis, typically including tasks such as handling missing values, outlier detection, normalization, and feature scaling. This step is crucial for ensuring accurate and reliable results in data analysis and modeling.

1) *Data transformation:* In the data transformation stage, several techniques are applied to prepare the data for modeling. Categorical variables are encoded into numerical representations, typically using methods like one-hot encoding to create binary columns for each category or label encoding to assign unique numerical values to categories. Numerical functions can be scaled to make sure consistent degrees across variables, assisting algorithms touchy to function magnitudes. Feature engineering consists of crafting new capabilities from

present ones, leveraging domain information or statistical insights to enhance model performance. This may consist of growing interaction terms, polynomial features, or transform variables to better capturing relationships or styles within the information. These transformation steps collectively intention to enhance the suitability and predictive strength of the dataset for subsequent modeling duties [22].

2) *Data exploration and visualization:* Explore the distribution of each feature and the relationship between feature variable and the targeted variable. Visualize the information using plots along with histograms, box plots, scatter plots, and so forth. to benefit insights into the records and become aware of patterns.

#### C. Automatic Medical Insurance Claims Service System through Blockchain Technology

The remedy offered by this study was to implement an autonomous insurance claim servicing system using the blockchain. The surroundings are used to exchange data between healthcare providers, insurance providers, and individuals. The environment's functions include the blockchain computing center (BCC), the appropriate government agencies (CA), the healthcare facility (MI), the insurance provider (IC), the finance company (BK), the patient (PT), & the center for arbitration (AI). Medicinal institutions can create a healthcare alliance chain under the supervision of the medicinal board CA1. Assurance and banks can join a financial alliances chain that is overseen by the banking regulator, CA2. Participants of the exact same alliance are able to exchange entire material.

Step 1: All CA, MI, IC, BK, and PT must verify with BCC in order to get both public and private ECDSA signing keys, as well as public and secret PKI key pairs. BCC also saves every patient's healthcare blockchain information. Furthermore, various kinds of CA will establish partnerships among the people they represent, and the partnership's membership' data will be exchanged.

Step 2: The patient, PT, buys health assurance through the health care firm IC. The IC will first check the PT's identification and then execute an insurance agreement with them. The PT must furnish the IC with the details of its BK account and paperwork will then sent to the BCC via the CA. Whenever the PT returns hospital in MI not too distant upcoming, and the examination result satisfies the alleged contented indicated in the health insurance agreement, the IC will move forward by the healthcare claims.

Step 3: Whenever a patient PT visits a healthcare facility MI and notifies the MI that they he or she has acquired health coverage, the MI will first authenticate the PT's identification, review the PT's electronic health record EMR, and then issue an authorization, with the information being communicated to the Sec via CA.

Step 4: The medical facility MI then notifies the assurance firm IC to process claims from insurance companies, and the IC acquires the PT medically-related diagnostic material given by MI.

Step 5: The insurance provider IC instructs the financial institution BK to pay the patient PT, and the record is transferred to the BCC via the CA.

Step 6: A claimed disagreement, the patient PT may file a complaint with the arbitration agency AI. AI will receive the communication contents from both side besides arrive at logical decisions.

#### D. Smart Contract Initialization

Blockchain technology was used in the suggested design. Certain essential data is kept and confirmed on the blockchain throughout the verification and permission procedure. The smart contract is a code that defines the block chain's most essential data. Everyone created essential data, which is stored on the blockchain in the suggested smart contract. Every smart contract has the following fundamental fields: id (identity), information about the transaction, certification, and timestamp. The smart contracts include the individual's bank account, whereas the smart contract includes the insurance company's bank account. The field's insurance contract is included with the smart contract. A smart contract supports digitized medical records. Finally, the purchase ID is shown in the smart contract. The blockchain technology center also provides both private and public key sets for every position during the authentication step.

1) *Registration phase:* The network's role X may include the Competent of authorities (CA), the healthcare facility (MI), the insurance provider (IC), the financial company (BK), and the individual in need (PT), who sign up blockchain center (BCC) also receive an individual's public/private key combination and a digital proof to verify their identities via a safe channel. Fig. 2 depicts the diagram for the enrollment process. Registration phase flow is explained in Fig. 2.

2) *Authentication process:* During the start of the interaction, system roles A and B must authenticate their respective identities using the ECDSA technique. System roles

A and B may comprise appropriate government agencies (CA), healthcare providers (MI), insurance firms (IC), banking (BK), and individuals (PT).

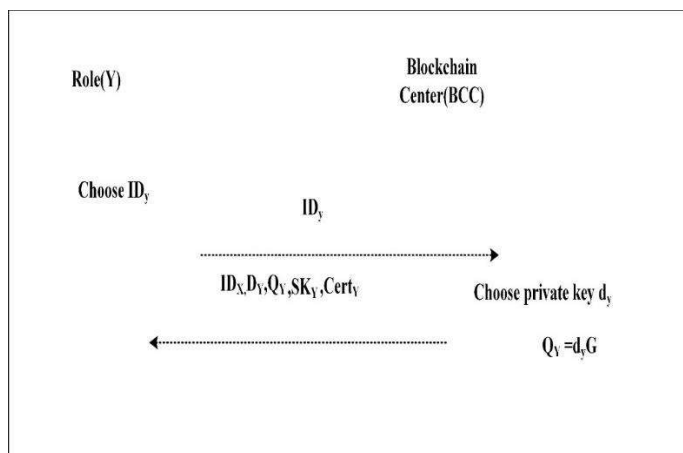


Fig. 2. Registration phase.

3) *Communications procedure:* The recommended solution makes use of the hyper ledger's block chain design, which increases the CA's role, allows for more versatility in accessing monitoring, and reduces the stress on BCC. After authenticating interactions across all roles, the details will then be provided to the various CAs, which will then send the blockchain information BCC. MI and IC both operate to own CA, which might allow documents flow throughout CA membership as well as cross-CA management of entry while retaining safety & efficacy. The accessible party (AP) might be a hospital (MI), an insurance company (IC), a financial institution (BK), or a client. A schematic representation of the CA communications method is proposed [15]. Blockchain-Based Medical Insurance System is shown in Fig. 3.

#### E. Event-Driven Architecture Insurance Claim Process

The Claim entity depicts an insurance claim which consumers can file or that current insurance companies may employ to assess how to pay out. The claim form includes a Loss Amount and a connection to the Insurance Policy organization. After an Event has been established, the processing Event method the request may be utilized to determine if it needs to be payed out. The Root Cause Mapping object, typically is a Boolean, is used within the process Event () method to determine if a payout is necessary for a fundamental issue and insurance policy combination. The Root Cause Mapping is defined in the privileges portion of the agreement.

1) *Identity NFT:* The identification NFT will serve as verification of identification for claim management. Whenever the program is first set up, the NFT is going to be coined (generated) for claims managers using the mailing addresses supplied in the initial setup script. Anyone is unable to establish an event if they have an Identification NFT.

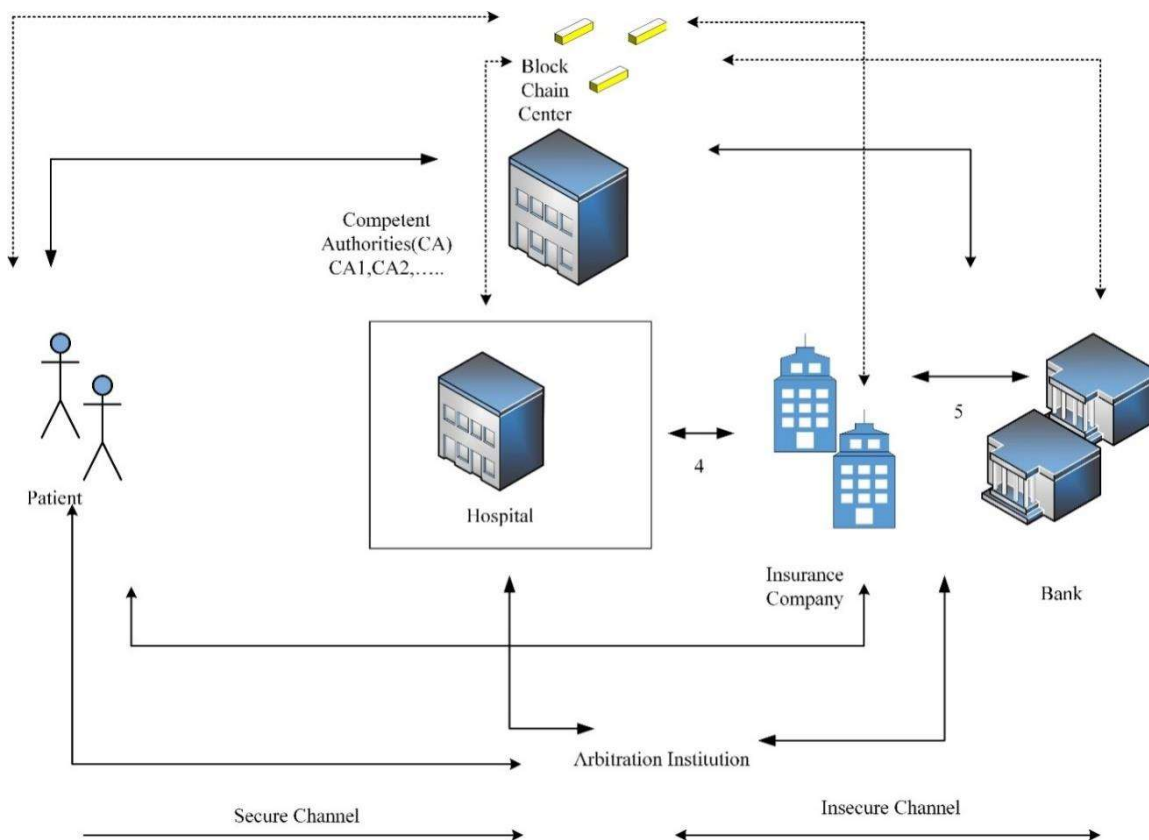


Fig. 3. Blockchain-based medical insurance system.

When the entitlements director organizes an event, their uniqueness is verified by the identification NFT contract. If it is valid, an event is produced in the Event Contract, which is then passed on to the claim contract for processing. All pending claims for the covered protocols are going to be adjusted using Root Cause Mapping. If the amalgamation of the root problem and regulation proves accurate, a payment should be provided. In the initial release of the program, just the claim's current state will be changed.

Any individual may file a right depending on their assurance coverage. The Entitlements agreement will compare information on the insurance against current claims. If previous demands exact similar policy and root cause were previously approved or postponed then the latest one will be assigned the identical status. In subsequent versions of the app, any blockchain-based insurance company may utilize this capability to poll whether or not an entitlement deserves to be paid out, allowing them to streamline this process.

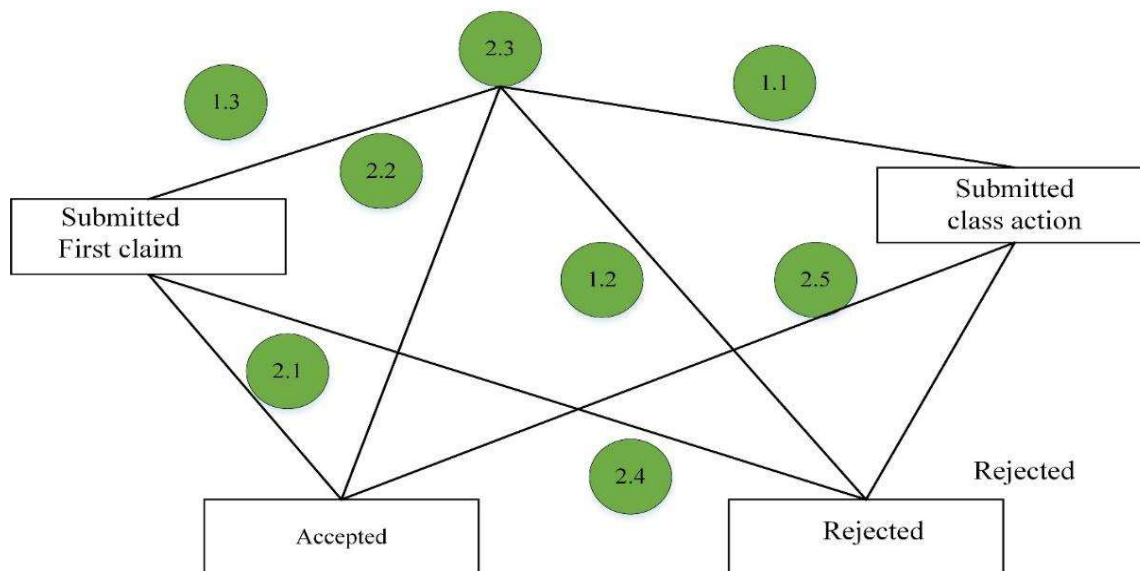


Fig. 4. State diagram of claims object.

The claim is the primary entity in the realm structure. The entitlement made in the framework can be in one of four "states" depending on the user's selections. Phase transitions 1.1 to 1.4 are for newly formed claims. Whenever the request is the initial one for a benefit, it will be marked given the status "submitted - first claim" (1.1). If more than one claim is currently lodged for the asset, it will be marked as "accepted - class action" (1.4). If a entitlements administrator has already accepted or rejected a claim having an identifiable cause, then subsequent claims will be immediately approved (1.2) or denied (1.3). Steps 2.1–2.4 apply to claims that were previously filed at the time the claims administrator reports an event. In that point, all filed claims are going to be immediately approved (2.1 and 2.3) or denied (2.2 or 2.4) [23]. Fig. 4 shows state diagram of claims object.

V. RESULT AND DISCUSSIONS

The integration of blockchain technology and smart contracts presents a transformative solution for the insurance industry, revolutionizing claims processing and payouts. By automating the entire process based on predefined events, such as natural disasters or medical emergencies, the proposed system eliminates manual claims submission, enhances efficiency, and ensures transparency and security. With payouts executed automatically upon event verification, bureaucratic hurdles are bypassed, leading to expedited processing times and reduced fraud risks. This innovative approach not only streamlines operations but also fosters trust and reliability, ultimately delivering significant benefits to insurers and policyholders alike in a future characterized by expedited, transparent, and trustworthy claims processing.

Event-driven architecture for insurance claim processes is given in Fig. 5 driving events such as policy updates, patient updates, and claim submissions trigger a series of actions. A graph depicting the number of driving events per second illustrates the system's real-time processing capabilities, enabling insurers to handle fluctuating workloads efficiently. This visualization aids in understanding system performance and scalability, ensuring timely and accurate processing of insurance claims.

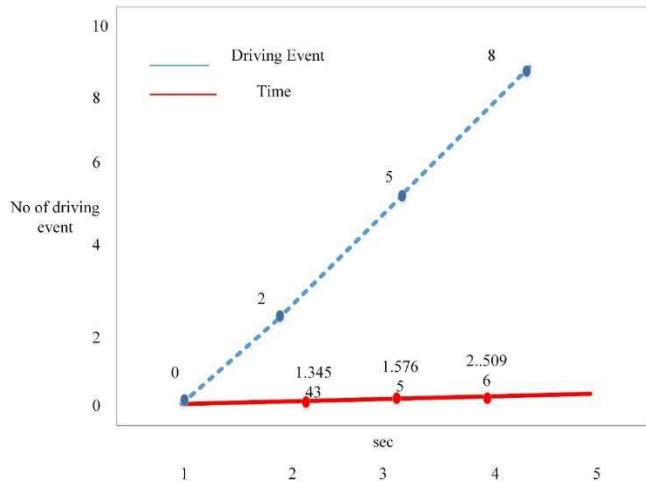


Fig. 5. Event-driven architecture for insurance claim processes.

Fig. 6 depicts a partial dependence plot illustrating how an old claim affects insurance outcomes. It visualizes the relationship between the age of a claim and its impact on insurance variables, such as claim probability or payout amount. This graph, insurers can understand how the age of a claim influences risk assessment and decision-making in insurance processes. It helps identify patterns and trends, enabling more informed underwriting and claims management strategies. This graphical representation facilitates data-driven insights for optimizing insurance operations and managing risk effectively.

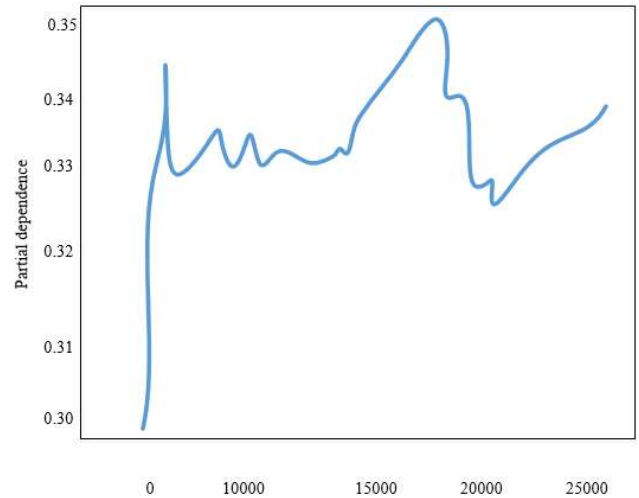


Fig. 6. Automatic insurance claim prediction.

Fig. 7 shows the amounts of paid and denied claims for different categories of old claims. The amount of paid claims is higher than denied ones, with the 3rd claim having the highest amount of paid claims. The bar chart helps to visualize the distribution and comparison of paid and denied claims for different categories of old claims.

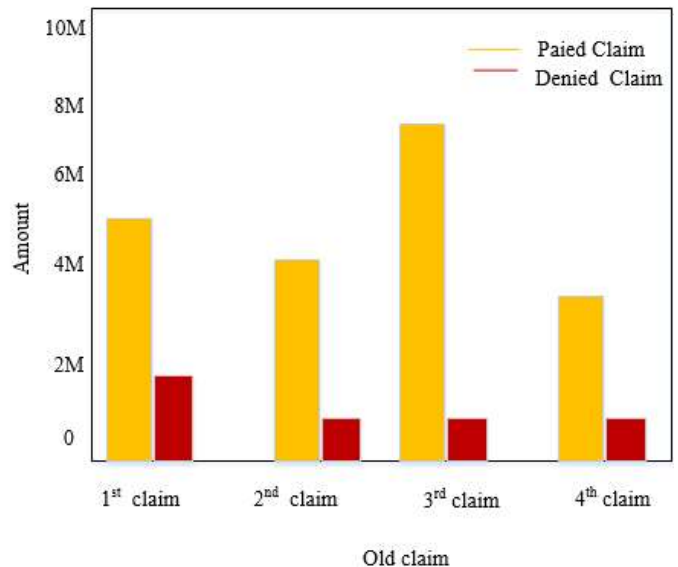


Fig. 7. The amounts of paid and denied claims.

Table I presents performance metrics for different methods the proposed method achieves high accuracy (94.44%) and outperforms others in precision (98.1%), recall (98.98%), and F1-score (98.54%) in Fig. 8. This indicates its effectiveness in correctly identifying positive instances while minimizing false positives and negatives, demonstrating its potential superiority in the classification task.

TABLE I. PERFORMANCE METRICS

Methods	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
RNN	90.54	90.4	92.00	92.44
Auto encoder	92.77	91.88	91.76	91.56
VAE	95.5	93.57	94.01	94.45
Proposed method	97.6	98.1	98.98	98.54

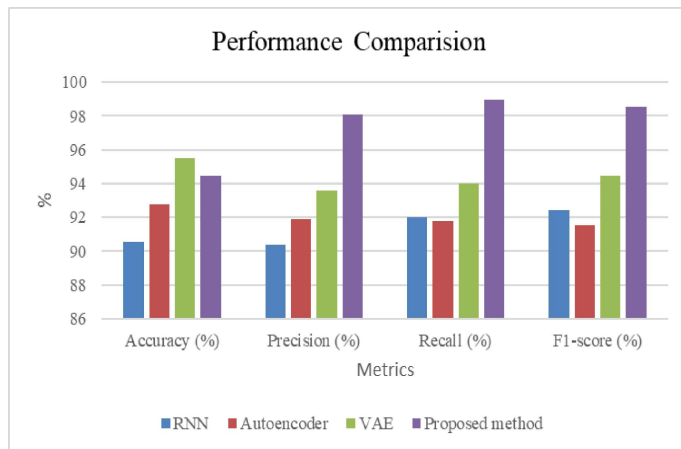


Fig. 8. Performance comparison.

Table II shows that the proposed method achieves an accuracy of 97.6% using the "Health Insurance Dataset - EDA". It also maintains high accuracy rates with other datasets: MedClaimsData (96.7%), HealthCoverStats (95.8%), and HealthinsureDB (93.7%).

TABLE II. DATASET COMPARISON

Dataset	Proposed Method Accuracy
Health Insurance Dataset-EDA	97.6%
MedClaimsData	96.7%
HealthCoverStats	95.8%
Healthinsure DB	93.7%

### A. Discussions

The proposed integration of blockchain technology and smart contracts in insurance claims processing offers several significant advantages. Firstly, it addresses the limitations of the existing method by adopting event-based smart contracts, eliminating the need for manual claims submission. This automation enables automatic triggers based on predefined events such as natural disasters or medical emergencies, accelerating the claims initiation process and ensuring accuracy and transparency. By doing so, it overcomes the existing challenges related to inefficiency and complexity,

which require physical visits to facilities for certification before submitting paperwork to insurance firms, leading to delays in compensation and posing challenges in data verification and risk management. By encoding specific conditions and triggers within smart contracts, the entire claims process becomes transparent, secure, and immutable, thereby minimizing the potential for fraud and dispute. Thirdly, the automatic execution of payouts upon verification of triggering events bypasses traditional bureaucratic procedures, leading to significantly reduced processing times [15]. Through the streamlined system outlined, insurers and policyholders stand to benefit from increased efficiency, transparency, and trustworthiness in claims processing, ultimately enhancing the overall insurance experience for all stakeholders. These advantages address the limitations of the existing method, such as challenges related to infrastructure development, legal obstacles, and data availability, thus making the proposed system a more robust and effective solution.

### VI. CONCLUSION AND FUTURE WORKS

The integration of blockchain technology and smart contracts has revolutionized the insurance industry by enhancing efficiency, transparency, and reliability. This research proposes a simplified system that automates the entire claims process from submission to reimbursement, eliminating the need for human claim filing. The system triggers policyholders' claims by predetermined occurrences, such as medical emergencies or natural disasters, allowing for prompt and precise claim initiation. Smart contracts, programmed with precise triggers and conditions, guarantees the transaction immutability, security, and transparency. Reimbursements are carried out automatically after the triggering event has been verified, reducing processing times and reducing fraud and disagreement. This innovative approach to insurance claims processing has shown significant reductions in processing time, minimized fraud potential, and enhanced transparency. The Python-implemented system achieved 97.6% accuracy, demonstrating its efficacy and reliability. This study contributes to addressing the limitations of existing insurance claim procedures by providing a streamlined, automated, and secure solution. By integrating blockchain technology and smart contracts, the insurance industry can overcome challenges of inefficiency, complexity, and lack of transparency in the current claims processing system. The research questions regarding the feasibility and effectiveness of event-based smart contracts in automating insurance claims processing have been successfully addressed, providing valuable insights for future implementations in the insurance sector.

### REFERENCES

- [1] "Chapter 7: SMART CONTRACTS in: FinTech." Accessed: Apr. 24, 2024. [Online]. Available: <https://www.elgaronline.com/edcollchap/edcoll/9781800375949/978180375949.00018.xml>
- [2] N. R. Bhamidipati et al., "Claimchain: Secure blockchain platform for handling insurance claims processing," in 2021 IEEE International Conference on Blockchain (Blockchain), IEEE, 2021, pp. 55–64.



- [3] C. Eckert, C. Neunsinger, and K. Osterrieder, "Managing customer satisfaction: digital applications for insurance companies," *Geneva Pap. Risk Insur.-Issues Pract.*, vol. 47, no. 3, pp. 569–602, 2022.
- [4] L. Zheng and L. Guo, "Application of big data technology in insurance innovation," in *International conference on education, economics and information management (ICEEIM 2019)*, Atlantis Press, 2020, pp. 285–294.
- [5] M. Hanafy and R. Ming, "Machine learning approaches for auto insurance big data," *Risks*, vol. 9, no. 2, p. 42, 2021.
- [6] L. Rukhsar, W. H. Bangyal, K. Nisar, and S. Nisar, "Prediction of insurance fraud detection using machine learning algorithms," *Mehran Univ. Res. J. Eng. Technol.*, vol. 41, no. 1, pp. 33–40, 2022.
- [7] D. E. Warren and M. E. Schweitzer, "When weak sanctioning systems work: Evidence from auto insurance industry fraud investigations," *Organ. Behav. Hum. Decis. Process.*, vol. 166, pp. 68–83, 2021.
- [8] J. Madir, "Smart contracts," in *FinTech*, Edward Elgar Publishing, 2021, pp. 175–198.
- [9] A. S. Mishra, "Study on blockchain-based healthcare insurance claim system," in *2021 Asian Conference on Innovation in Technology (ASIANCON)*, IEEE, 2021, pp. 1–4.
- [10] V. Kalsgonda and R. Kulkarni, "Role of Blockchain Smart Contract in Insurance Industry," Available SSRN 4023268, 2022.
- [11] X. Lin and W. J. Kwon, "Application of parametric insurance in principle-compliant and innovative ways," *Risk Manag. Insur. Rev.*, vol. 23, no. 2, pp. 121–150, 2020.
- [12] K. L. Narayanan, C. R. S. Ram, M. Subramanian, R. S. Krishnan, and Y. H. Robinson, "IoT based smart accident detection & insurance claiming system," in *2021 Third international conference on intelligent communication technologies and virtual mobile networks (ICICV)*, IEEE, 2021, pp. 306–311.
- [13] J. C. Mendoza-Tello, T. Mendoza-Tello, and H. Mora, "Blockchain as a healthcare insurance fraud detection tool," in *Research and Innovation Forum 2020: Disruptive Technologies in Times of Change*, Springer, 2021, pp. 545–552.
- [14] A. Borselli, *Smart contracts in insurance: a law and futurology perspective*. Springer, 2020.
- [15] C.-L. Chen, Y.-Y. Deng, W.-J. Tsaur, C.-T. Li, C.-C. Lee, and C.-M. Wu, "A traceable online insurance claims system based on blockchain and smart contract technology," *Sustainability*, vol. 13, no. 16, p. 9386, 2021.
- [16] D. Oza, D. Padhiyar, V. Doshi, and S. Patil, "Insurance claim processing using RPA along with chatbot," in *Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST)*, 2020.
- [17] N. Dhieb, H. Ghazzai, H. Besbes, and Y. Massoud, "A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement," *IEEE Access*, vol. 8, pp. 58546–58558, 2020.
- [18] F. Loukil, K. Boukadi, R. Hussain, and M. Abed, "Ciosy: A collaborative blockchain-based insurance system," *Electronics*, vol. 10, no. 11, p. 1343, 2021.
- [19] N. Fernando, A. Kumarage, V. Thiyaganathan, R. Hillary, and L. Abeywardhana, "Automated vehicle insurance claims processing using computer vision, natural language processing," in *2022 22nd International Conference on Advances in ICT for Emerging Regions (ICTer)*, IEEE, 2022, pp. 124–129.
- [20] K. Dutta, S. Chandra, M. K. Gourisaria, and G. Harshvardhan, "A data mining based target regression-oriented approach to modelling of health insurance claims," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, 2021, pp. 1168–1175.
- [21] "Health Insurance Dataset - EDA | Kaggle." Accessed: Feb. 15, 2024. [Online]. Available: <https://www.kaggle.com/code/mregoyau/health-insurance-dataset-eda>
- [22] S. Manikandan, "Data transformation," *J. Pharmacol. Pharmacother.*, vol. 1, no. 2, p. 126, 2010.
- [23] S. Gillis, "Blockchain-based Application for Insurance Claims Management," PhD Thesis, Harvard University, 2023.



## Student Learning Based Data Science Assisted Recommendation System to Enhance Educational Institution Performance

D. Naga Jyothi<sup>1\*</sup>

Uma N Dulhare<sup>2</sup>

<sup>1</sup>*Chaitanya Bharathi Institute of Technology, Hyderabad, Telangana, India*

<sup>2</sup>*Muffakam Jah College of Engg. And Tech, Hyderabad, Telangana, India*

\* Corresponding author's Email: [dnagajyothi\\_cseaiml@cbit.ac.in](mailto:dnagajyothi_cseaiml@cbit.ac.in)

---

**Abstract:** The student feedback data possesses to be the fundamental influencers of decision-making process in diverse applications. The performance prediction based on student's feedback about the educational institution helps for better solution recommendation. The automated solution recommendation based on student's feedback extensively support the educational institution to make better decisions for improvisation. In most of the existing research works, the performance can be analysed, but suitable solution recommendation is not provided. Also, the existing recommendation works fail to generate accurate outcomes, consumes more time with higher error rates. Hence on diminishing the existing issues, this research work presents a Data science-based solution Recommendation model based on hybrid deep learning approaches. Pre-processing, feature extraction, feature clustering, performance prediction, and recommendation are the steps in the suggested model. In this research, the student feedback data is collected from Kaggle source and some of the attributes are added manually. Pre-processing techniques for the text data include stop-word-removal, tokenization, case-folding, and stemming. The features are extracted using Enhanced Lexicon bidirectional encoder representations from transformers (ELexBert) model. The significant attributes are selected using Adaptive Coati optimization (ACoaT) algorithm. The selected features are clustered based on feature similarity using Upgraded density based k-means clustering (Uden\_KMC) model. A new type of hybrid deep learning model known as Channel Block Densnet with Dilated Convolution BiLSTM (ChaBD-BiL) is employed for recommending better decisions. The recommendation performances using feedback data are evaluated using PYTHON where the overall accuracy of 98.19%, specificity of 98.25%, F1 score of 91.28%, sensitivity of 97.41% and Kappa score of 91.28% are obtained.

**Keywords:** Student feedback, Lexicon BERT, Coati optimization, Density clustering, Channel block DenseNet, Dilated convolution.

---

### 1. Introduction

Due to the development of immense database and information technologies, the data science holds a greater impact for promoting the progress of data analysis [1]. Diverse studies insists that the applications of data science can be categorized into several technologies like machine learning, deep learning, and ensemble approaches. The education system is computerized nowadays to render better education to the students [2, 3]. The data science can effectively manage the requirements of students, gather better knowledge, make better decisions and

also examine the performance of the educational Institution in a great way. The key role of Educational Data Mining is to discover and overcome the research issues in the field of education [4]. The collection of student feedback related to teaching and other learning activities can assist for the investigation [5]. Through the optimal investigation process, the opportunities, strengths, threats, and weakness in the education system can be analysed properly and further actions can be taken.

Effective recommendations can be provided to educational institutions to support the students in enhancing their studies and assisting the instructors

to improve their teaching effectiveness [6]. Due to lack of relevant information, the educational institutions are suffering extensively to support the students by resolving the issues [7]. The only way to gather the information from students is getting direct feedback from the students. One of the traditional feedback mechanisms to gather data are filling of forms directly by the students [8]. Those mechanisms possess huge number of issues like only a certain question set are provided to the students [9, 10]. The chances to expose other forms of issues related to the educational system are not provided to the students. As the traditional mechanism is highly time consuming, online feedback mechanism is pursued in most of the institutions.

The online feedback mechanism is highly significant to gather student's feedback based on different attributes [11]. The students can give the suggestions to the educational institutions more effectively using online feedback mechanism compared to the traditional mechanism. The faculty members are supposed to utilize the feedback to determine their strengths and areas of improvement [12, 13]. Even though online feedback mechanism serves to be better, the analysis of each feedback and appropriate actions to be taken are highly complex [14]. There is so much of research carried out previously on processing the student feedback data. But in most of the works, only sentiments associated to the input data are analysed whereas appropriate recommendation is still lagging [15]. The recommendation system for enhancing the performance of educational institution have attracted huge attention for enhancing the student performance.

To overcome the challenging issues and to recommend better suggestions to the educational institution, an automated recommendation system based on effective feedback prediction is highly required [16]. Many recommendation-based research is carried for improving the student's future based on the student details [17]. However, suggestions for improving the functioning of educational institutions based on student feedback are quite innovative. Numerous models based on machine learning (ML) and Artificial Intelligence (AI) are used for promoting effective recommendation [18, 19]. But more time complexity, inefficiency in generating precise outcomes, increased rates of error and degraded training ability are found to be the challenging issues. Different software solutions are built utilising familiar programming languages to make it easier for designers to use machine learning technology and

predictive analytics. [32]. Recently, deep learning (DL) [20] based models are widely used as it produces faster and precise outcomes.

## 2. Motivation

The educational data science insists the use of data collected from educational environments for overcoming the issues through suitable decisions. Data science is a concept employed to merge the analysis of data, statistics, and feedback by using effective technologies. Various algorithms which addressed the classification problems are evaluated within the education science sector [31]. Algorithms for optimisation can be classified as probabilistic or deterministic which can be used for selection of the optimized features [34]. In the educational institution, the feedback of students is highly necessary for improving the performance further. In the recent days, computerization process is widely used by the educational institutions tending to the creation of huge amount of data. The collected feedback data from the students would be highly helpful for the teachers, administrators and so on for better decision making. Anyhow based on the student's feedback, recommendation of appropriate solutions to the educational institution is highly challenging and consumes more time. The existing research highly concentrates on data processing and categorizing the input texts based on sentiments like positive, negative and neutral, but effective solution recommendations are not performed. Also, the outcomes cannot be predicted much accurately and if predicted also, often results in increased rates of error. There is lack of research performing recommendation of suitable solutions based on student's feedback to enhance the educational institution performance. Due to ineffective consideration of student's feedback, the performance of educational institution as well as the students are influenced. Hence, an automated recommendation model is highly required to fulfil the student's requirement. Motivated by the existing challenges, data science assisted hybrid Deep Learning model is presented in the suggested study to obtain enhanced recommendation solutions.

The following lists some of the major contributions made by the suggested model:

To extract the effective features using Enhanced Lexicon bidirectional encoder representations from transformers (ELexBert) model and choose the best features utilising Adaptive Coati optimization (ACoaT) algorithm.

To generate clusters using Upgraded density based k-means clustering (Uden\_KMC) by considering the similarity of features.

To introduce a data science-based solution recommendation model using Channel Block densenet with Dilated Convolution BiLSTM (ChaBD-BiL) network with enhanced accuracy and less rates of error.

To utilize Channel Block densenet for prediction and Dilated Convolution BiLSTM for recommending a suitable solution.

The suggested method's higher performance would be demonstrated by assessing its performances with the current state-of-the-art approaches using several performance indicators.

The suggested research work is well structured into different sections. In Section-2, a few prediction and recommendation works conducted by different researchers are surveyed. The new approaches to text processing are shown in Section-3 to explain how the recommended methodology operates. In Section-4, the models used to analyse the performance are covered. The suggested research work's Conclusion and Future scope is presented in Section-5 along with the appropriate references.

### 3. Related works

Some of the recent prediction and recommendation works in text processing are specified as follows.

Karaoglan Yilmaz, Fatma Gizem, and Ramazan Yilmaz [21] investigated the opinions of aspiring educators about personalised recommendations based on learning analytics. Based on the flipped learning model, the research was undertaken on 40 teachers in computer course. The outcomes of learning analytics were obtained based on user activity in the learning management system (LMS) of students. Semi-structured opinion surveys were used to gather research data, and content analysis was done based on that data. The effective aspects and demerits of guidance feedback and personalized recommendation dependent upon the learning analytics can be analysed through this research. The research says student-centric learning analytics can be considered, and student opinions can be evaluated for decision making process. Recommendations can be provided to the students to enhance the metacognitive thinking skills.

Sood, Sakshi, and Munish Saini [22] utilized an integrated approach comprising of Cluster-based Linear Discriminant Analysis (CLDA) and Artificial Neural Network (ANN). The major focus of this research was to recommend the motivational

comments to the probable students. As a result, students can choose relevant courses, and the suggested remarks help students understand why they may have dropped out. Through this research, the number of dropouts can be extensively minimized with the suitable selection of courses to enhance the overall performance. One benchmark and one synthetic dataset were used, and they are pre-processed initially to process this research. This research talks about the usage of IoT with the wearable devices in the next works to collect the real-time data and to compare the student performance which can reduce the student dropouts.

Yangsheng, Zhang [23] constructed an intelligent model for sports evaluation with the integration of AI based teaching system based on neural network modelling. The final evaluation and process determination are where the AI model starts. The recurrent neural network (RNN) was employed for data analysis and training. In addition, a new decoder was established to process data and a simplified gated neural network (GNN) was developed to construct the internal model structure. In accordance with this, a control experiment was designed to examine the model execution. Through this research, a better outcome can be obtained in predicting the performance by considering the sports students. This research also says about the usage of enhanced neural network and AI based algorithms in future for student performance prediction with better analysis and accuracy.

Kanetaki, Zoe, Constantinos Stergiou, Georgios Bekas, Christos Troussas, and Cleo Sgouropoulou [24] explored the prediction of grades in online engineering education. After being eliminated from statistical analysis, a hybrid model with 35 variables was created and found to have a good correlation with students' academic achievement. Initially, a Generalized Linear- Model was involved and later its errors were employed as an additional related variable to the Artificial Neural Network (ANN). This research predicts that grade as a dependant variable can be a best variable for success of the model. The survey answers of 158 students were validated in this work by dividing the dataset into three subsets. The particulars like standard error, p-value and coefficients were estimated for all variables. The future work of the research talks about the model performance prediction for the next batch of students. A confusion matrix can be used, statistical significance can be found for the variables and model accuracy can be tested for that batch of students.

Ouyang, Fan, Mian Wu, Luyi Zheng, Liyin Zhang, and Pengcheng Jiao [25] combined AI based

performance prediction approaches and learning analytic methods to boost the learning effects of students in Collaborative learning context. The major purpose of this research was to show the predicted outcomes to students as well as course instructors which can improve the learning quality and teaching performance. It has shown the differentiations of collaborative learning effect over students with and without the integrated approach. The quasi-experimental research was carried on the online engineering courses. Effective enhancement of students, enhanced performances of collaborative learning and student satisfaction strengthening were the outcomes analysed in this research. The future work of this research should use the expanded educational contexts with an increased sample test set . It mainly suggests to propose a integrated approach of AI and LA, conduct the statistical studies using the integrated approach to provide a clear path between AI and Education Domain.

Kusuma, Purba Daru, and Ashri Dinimaharawati [39] proposed The extended stochastic coati optimizer (ESCO), a new metaheuristic, is presented in this paper. The flaw in the coati optimisation algorithm (COA) is expanded to create ESCO. The amount of searches and references included in COA is increased by ESCO. This research work has helped to get a good understanding of Coati Algorithm and its extended version which splits the population into two fixed groups, each performing its strategy for feature optimisation.

To conquer the future works and drawbacks faced in the existing algorithms, a novel hybrid DL model is presented to promote effective recommendation solutions based on the input student feedback data. The procedure of proposed methodology has been provided step by step as follows.

#### 4. Proposed methodology

The student feedback holds to be the fundamental influencers of decision-making process. The performance prediction of student's feedback about the educational institution helps for better solution recommendation. The automated solution recommendation based on student's feedback extensively support the educational institution to make better decisions for improvisation. In most of the existing research works, the performance can be analysed but suitable solution recommendation is not provided. Also, the existing recommendation works fail to generate accurate outcomes, consumes more time with higher error rates. Hence on diminishing the existing issues, this research work

presents a data science-based solution recommendation model based on hybrid deep learning approaches. Fig. 1 explains the schematic representation of suggested workflow.

The student feedback data is collected from online Kaggle source. Additionally, some of the attributes and recommendation solution are manually added in the dataset to process this research work. The steps involved in student feedback-based solution recommendation model are listed as follows.

- Pre-processing
- Feature - Extraction
- Feature - selection
- Feature - Clustering
- Performance - prediction
- Recommendation

Initially, pre-processing is carried using stemming, tokenization, case folding and stop word removal. The characteristics are extracted from the pre-processed data using ELexBert model. The most relevant features are selected using ACoaT algorithm. The selected features are clustered into diverse groups based on feature similarity using Uden\_KMC model. From the generated clusters, the performances of educational institution based on student feedback are predicted and suitable solutions can be recommended based on that prediction. This can be performed using a novel hybrid DL model called ChaBD-BiL. Here, Channel Block dense net is used for prediction where the educational feedback given by the student can be predicted as good, not bad and poor. Based on the predicted outcomes, Dilated Convolution BiLSTM recommends for a suitable solution to overcome the issues.

#### 4.1 Text pre-processing

The gathered input text data is subjected to abundant irrelevant data that highly declines the quality of text and overall system performance. To attain enhanced performance, significant input text data is necessary and so text data pre-processing [26] is initially carried out in the proposed DL model. Through pre-processing, structured text data can be attained that is significantly crucial for precise recommendation system. In the proposed recommendation model, steps like stemming, tokenization, case folding and stop word removal are used for pre-processing the text data. The explanation of every pre-processing step undertaken are clearly described as follows.

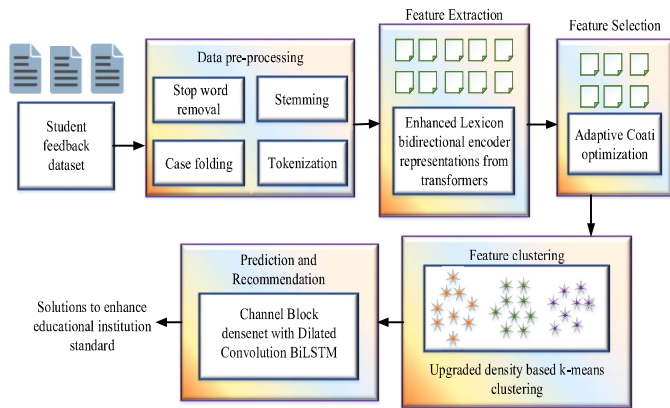


Figure. 1 Block architecture of proposed model

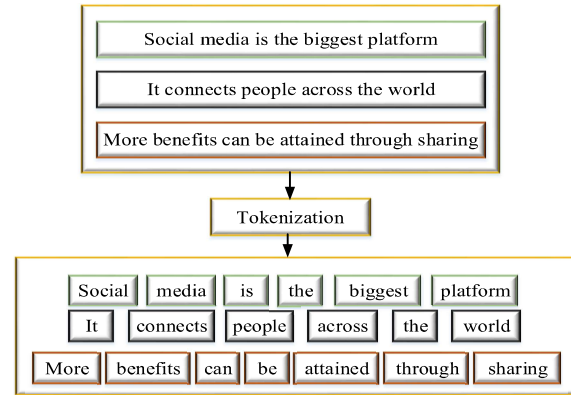


Figure. 2 Instance of Tokenization process

#### 4.1.1. Case folding

Case folding process is performed in the proposed model to convert the letters from text documents to corresponding lower or upper case. In text pre-processing, case folding has been utilized to convert letters into lowercase format.

#### 4.1.2. Tokenization

The process of tokenization is considered as one of the most effective tasks in text data processing. The process of separating a sentence, paragraph, entire text or phrase into small units or words is called as tokenization. The smaller units separated from text data are said to be tokens. In text language processing, the words that determine character string must be recognised and so tokenization process acts as a significant step. An instance of tokenization process performed in the text data are insisted in Fig 2.

#### 4.1.3. Stop word removal

Removal of stop words from text data tends to be a crucial process that is undertaken during pre-processing stage. The major objective of stop word eradication is to remove the words that are usually found through the textual data. Essentially in pronouns, English verdicts, articles, and prepositions present in the given data are considered to be stop words. In text mining-based applications, stop word removal is carried out for analysing relevant words. An example for stop - word - removal for the given text input data is established in Table 1.

#### 4.1.4. Stemming

The process of producing morphological variant of base word is known as stemming. Stemming assists in reducing a word over its corresponding

Table 1. An example for stop – word - removal

A text sample with stop words	Text after stop word removal
He wishes to eat an apple	“Wish”, “Eat”, “Apple”
The dress appears very pretty	“Dress”, “Appear”, “Pretty”
How to deliver a book in office	“Deliver”, “Book”, “Office”
The woman brings bag on her hands	“Woman”, “Bring”, “Bag”, “Hand”

Table 2. An instance for Stemming

Sample word	After Stemming
Connecting	Connect
Introducing	Introduce
Call	Call
Building	Build

word stem that merges the root words. An instance of stemming dependent upon the sample word is given in Table 2.

### 4.2 Feature extraction

Feature extraction is the process of identifying important features from pre-processed text material to improve performance overall. The principal objective of feature extraction is identification of relevant features for enhancing the recommendation efficiency. Various Deep Learning models can be applied for Feature extraction [38]. In DL model, BERT [27] is considered as one of the significant word embeddings and it can efficiently learn the word contexts. To enhance the efficiency of BERT model further, ELexBert is employed in the recommended model. The design idea of ELexBert model is to utilize Lexicon selected N-grams, convert lexicons into vectors and apply BERT embedding algorithm to obtain a relevant set of

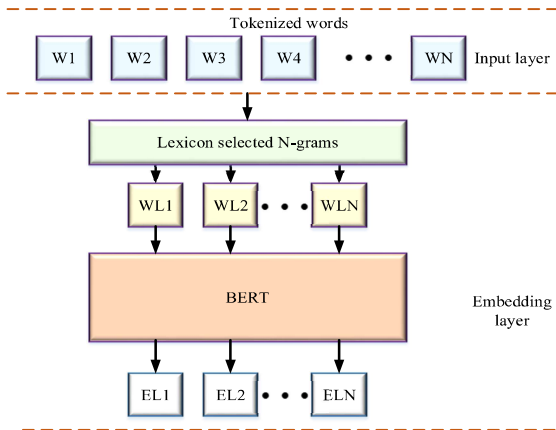


Figure. 3 ELexBert model representation

features. The architecture of the proposed ELexBert model can be seen in Figure 3.

The N-grams denote the combination of words from a sentence that generates a markovian process and is used to determine the subsequent word in a string of words. Also, it generates co-occurrence of words from text in a more significant manner. For instance, the N-grams considered from the sentence is given as follows.

$$\text{Sentence} = \{W1, W2, W3, \dots, WN\} \quad (1)$$

From the above expression,  $WN$  denotes the number of N-gram words. For diverse values of  $N$  (uni - gram, bi - gram), the set of lexicon selected N-grams get varied. For example,

$$\text{For } N = 1, N1 = \{W1, W2, W3, \dots, WN\} \quad (2)$$

$$\text{For } N = 2, N2 = \{W1\_W2, W2\_W3, W3\_W4, \dots, WN - 1\_WN\} \quad (3)$$

With the utilization of N-grams, it is applicable to choose a section from overall input text. This condition guarantees that the relevant words can be utilized when constructing the text vectors. Every word is converted into vectors using Lexicon to vector approach. To obtain better representation of vectors, a transformer structure is used by the BERT model to acquire contextual knowledge. The model makes extensive use of a multi-headed self-attentive mechanism to mine the data.

### 4.3 Feature selection

Feature selection is the process of eliminating duplicate and unnecessary information from a dataset using evaluation criteria to increase accuracy [35]. The use of metaheuristic algorithms has been crucial in the solving of complex issues [36]. The higher dimensionality features may tend to

maximize the computational complexity whereas precise outcomes cannot be obtained. Hence from the extracted features, the most optimal features of diverse text attributes are chosen using ACoaT algorithm.

Each process in the algorithm is described with a detailed formalisation in Eq. (4) to Eq. (14). The list of annotations used in this model are shown below.

$Z_p$	position of the feature in search space
$Z_{p,q}$	value of the feature
$X$	Number of features
$F$	objective function
<b>Iguana</b>	search space position of iguana
<b>Low , Upp</b>	lower and upper bounds of the decision variable
$T$	iteration counter
$T_c$	tent chaotic map
$t$	maximum number of iterations

The Coati optimization algorithm [28] is developed on analysing diverse coati behaviours. The coati positions or the features are initialized randomly using the below expression.

$$Z_p: z_{p,q} = \text{Low}_q + \text{Random}(\text{Upp}_q - \text{Low}_q), \quad \text{where } p = 1,2,3, \dots, X, \quad q = 1,2,3, \dots, \quad (4)$$

Here,  $Z_p$  indicates the position of  $p^{\text{th}}$  feature in search space,  $z_{p,q}$  symbolises the value of  $q^{\text{th}}$  variable and  $X$  specifies the number of features.  $\text{Low}_q$  and  $\text{Upp}_q$  signifies the lower and upper bound of  $q^{\text{th}}$  decision variable.  $\text{Random}$  represents the random real number between the range 0 to 1.

The strategy of attacking and hunting iguanas

The initial stage, known as the exploration phase, modifies the search space's properties while using the fitness function of minimised error rate. The place of best solution among the features is considered as the iguana position. According to popular belief, some coatis climb the tree while others wait for the iguana to fall. The following can be used to indicate how coati's position is updated at each iteration.

$$Z_p^{t+1}: z_{p,q}^{t+1} = z_{p,q}^t + \text{Random}(\text{Iguana}_q - \delta \cdot z_{p,q}^t), \quad \text{where } p = 1,2, \dots, [X/2], \quad q = 1,2, \dots, Y \quad (5)$$

From the above expression,  $\delta$  specifies the integer chosen randomly as equal to 1 or 2. The iguana is placed in an arbitrary location inside the

search area as it hits the ground. The coatis move and are replicated in the following expressions based on it.

$$\begin{aligned} \text{Iguana}_G: \text{Iguana}_G^q &= \\ \text{Low}_q + \text{Random}(\text{Upp}_q - \text{Low}_q), \\ \text{Where } q &= 1,2,3,\dots Y \end{aligned} \quad (6)$$

$$\begin{cases} Z_p^{T1}: z_{p,q}^{T1} = \\ \begin{cases} z_{p,q} + \text{Random}(\text{Iguana}_G^q - \delta \cdot z_{p,q}), \\ \quad F_{\text{Iguana}_G} < F_p \\ z_{p,q} + \text{Random}(z_{p,q} - \text{Iguana}_G^q), \\ \quad \text{else,} \end{cases} \\ \text{for } p = \lfloor \frac{X}{2} \rfloor + 1, \lfloor \frac{X}{2} \rfloor + 2, \\ \dots X \text{ and } q = 1,2, \dots Y \end{cases} \quad (7)$$

If the new coati position meets the fitness function, it can be updated at a reasonable cost; if not, the original position is retained. The revised strategy for  $p = 1,2,\dots X$  is simulated using the below given expression.

$$Z_p = \begin{cases} Z_p^{T1}, & F_p^{T1} < F_p \\ Z_p, & \text{else} \end{cases} \quad (8)$$

From the above expressions,  $Z_p^{T1}$  denotes the new position estimated for  $p^{\text{th}}$  coati,  $Z_{p,q}^{T1}$  denotes its  $q^{\text{th}}$  dimension,  $F_p^{T1}$  indicates the objective function and Iguana indicates the search space position of iguana.  $\text{Iguana}_G$  shows the iguana position on ground.  $F_{\text{Iguana}_G}$  indicates the objective function value,  $[\cdot]$  represents the greatest integer function.

The technique of escaping from predators

The animal flees from its place during the exploitation phase when it is attacked by a predator. In order to replicate the updating behaviour, a random position is generated in close proximity to the current coati location, as stated below.

$$\begin{aligned} \text{Low}_q^{\text{Local}} &= \frac{\text{Low}_q}{T}, \text{Upp}_q^{\text{Local}} = \frac{\text{Upp}_q}{T}, \\ \text{Where } T &= 1,2,3, \dots t \end{aligned} \quad (9)$$

$$\begin{aligned} Z_p^{T2}: z_p^{T2} &= z_{p,q} + (1 - 2\text{random}). \\ &\left( \begin{array}{c} \text{Low}_p^{\text{Local}} + \\ \text{Random}(\text{Upp}_p^{\text{Local}} - \text{Low}_p^{\text{Local}}) \end{array} \right), \\ \text{where } p &= 1,2, \dots, X, q = 1,2, \dots, Y \end{aligned} \quad (10)$$

The newly estimated point is adequate if it enhances the actual function value and the

requirement simulates using the below given expression.

$$Z_p = \begin{cases} Z_p^{T2}, & F_p^{T2} < F_p \\ Z_p, & \text{else} \end{cases} \quad (11)$$

The new position estimated for  $p^{\text{th}}$  coati based on exploitation phase is denoted as  $Z_p^{T2}$ . The  $q^{\text{th}}$  dimension is denoted as  $Z_{p,q}^{T2}$ ,  $F_p^{T2}$  denotes the objective function value, T indicates the iteration counter,  $\text{Low}_q^{\text{Local}}$  and  $\text{Upp}_q^{\text{Local}}$  represents the lower and upper bound of  $q^{\text{th}}$  decision variable. The ACoaT algorithm is utilised to improve the efficiency of selection performance. Tent chaotic map is used in the initialization strategy to swap random generation, and equation (4) can be rephrased as follows:

$$\begin{aligned} Z_p: z_{p,q} &= \text{Low}_q + T_c(\text{Upp}_q - \text{Low}_q), \\ \text{where } p &= 1,2,3, \dots, X, \quad q = 1,2,3, \dots, Y \end{aligned} \quad (12)$$

$$T_c^{t+1} = \begin{cases} \frac{T_c^t}{k}, & \text{Tent}^t \in (0, k) \\ \frac{1-T_c^t}{1-k}, & \text{Tent}^t \in (k, 1) \end{cases} \quad (13)$$

The coati position is adjusted using  $T_c$  tent chaotic map that assists to enhance the global searching performance. During the attack phase, the coati position is updated by the dynamic weight factor  $\rho$ . At the iteration end,  $\rho$  reduces adaptively where the coati performs a well local searching by maximizing the speed of convergence. Equation (7) can be reframed as below.

$$\begin{cases} Z_p^{T1}: z_{p,q}^{T1} = \\ \begin{cases} z_{p,q} + \text{Random}(\text{Iguana}_G^q - \delta \cdot z_{p,q}), \\ \quad F_{\text{Iguana}_G} < F_p \\ z_{p,q} + \text{Random}(z_{p,q} - \text{Iguana}_G^q), \text{ else} \end{cases} \end{cases} \quad (14)$$

The iteration counter and the maximum number of iterations are represented by the expression above. By selecting only, the most relevant features for prediction, this technique helps to solve the dimensionality problems by identifying the best features. One of the research works reiterates that, contrary to many other metaheuristics, interacting with as many individuals as possible has been shown to be more effective than doing so with only a limited group of people [40].



#### 4.4 Grouping of features

The selected features are clustered into diverse groups based on feature similarity using Uden\_KMC model. The K-means Clustering Algorithm (KCM) [29] is a separation-based cluster analysis approach. The initial step of KCM is to choose R number of objects or features as primary cluster centres. Assign each data point to the cluster associated with the nearest centre. The average of all data points assigned to each centroid is calculated. This average becomes the new centroid for the cluster. Each centroid is moved to the mean of its associated data points. The process is repeatedly carried out until a better convergence is accomplished. The KCM is extremely delicate over principal cluster centres and hence the clustering results vary based on principal cluster centres. This influences the mean point valuation, diverges the cluster center and so declines the clustering result. Hence, Uden\_KMC approach is employed for clustering in the proposed method.

In density based outlier detection, k- nearest neighbour (knn) distance and k-neighbourhood of every object is created primarily by Local Outlier factor (LOF). The distance between every object in its k-neighbourhood is estimated. Finally, the local outliers are identified by LOF and the outlier detection process is given as follows. The list of annotations used in Eq. (15) to Eq. (19) are shown below.

- u object
- (u,v) KNN distance
- n Number of Features
- Lde Local Density Estimator
- LOF Local Outlier Factor
- r Number of Features as primary cluster centers
- R number of objects or features as primary cluster centres

Step 1: Estimate the knn distance as  $(u, v)(v \in N_k(u))$  of every object  $u$ . The distance  $(u, v)$  is expressed as the straight distance connexion between objects  $u$  and  $v$ ,

$$\text{Distance} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \dots + (a_n - b_n)^2} \quad (15)$$

In the above expression,  $n$  represents the number of features.

Step 2: Assess the object density of  $u$  and it replicates the neighbourhood data distribution represented as the reciprocal of knn mean. The knn

local density of  $u$  is indicated below. ‘ $r$ ’ is the number of features as primary cluster centres.

$$\text{Lde}(u) = \frac{1}{\frac{1}{r} \sum_{s=1}^r \text{Dist}(u,v)} \quad (16)$$

Where Lde is Local density estimator.

Step 3: Estimate the LOF value of  $u$ .

$$\text{LOF}(u) = \frac{\sum_{s=1}^r \frac{\text{Lde}(v)}{\text{Lde}(u)}}{r} \quad (17)$$

Where LOF is Local Outlier Factor

The knn local density of  $u$  is indicated as  $\text{Lde}(u)$  and  $\text{LOF}(u)$  replicates the extent of  $u$  as an outlier. If  $\text{LOF}(u)$  is particularly greater than 1,  $u$  subjects to be isolated and so the object is not considered. The generated features are clustered using Uden\_KMC model and the procedure is listed as follows. The features to be clustered are  $C\{f_1, f_2, \dots, f_n\}$  and the output is to accomplish ‘ $n$ ’ number of clusters. In Uden\_KMC model,  $\text{LOF}(u)$  is evaluated using equation (17) and if  $\text{LOF}(u)$  value is greater than one, the isolated points are eradicated. The mean of  $F$  features is estimated as the first cluster center which is given as follows.

$$F_1 = \frac{1}{r} \sum_{s=1}^r W_s \quad (18)$$

Evaluate the following cluster center and then assess the distance between cluster center and residual points using the below expression.

$$M_r = \sum_{l=1}^R \text{Max}(z_{k-1}^l - \|W_r - W_l\|^2, 0) \quad (19)$$

From the above expression,  $W_r$  indicates the sample point whose  $M_r$  is the largest upcoming cluster center. Assess the distance between every  $W_s$  object, cluster center and allocate to the nearby cluster. Repeat the distance and mean calculation until active convergence is accomplished. Through Uden\_KMC model, the isolated feature points are lost from the data and the similar features of student review data can be grouped into diverse clusters.

#### 4.5 Recommendation model for better solutions

From the generated clusters, the performances of educational institution based on student feedback are predicted and suitable solutions can be recommended based on the prediction. This can be performed using a novel hybrid DL model called ChaBD-BiL. Here, Channel Block densenet is used for prediction whereas the educational feedback given by the student can be predicted as good, not

bad, and poor. Based on the predicted outcomes, Dilated Convolution BiLSTM recommends for a suitable solution to overcome the issues. The BiLSTM model effectively addresses the issues of parameter count and data stability [37]. Fig. 4 describes the schematic representation of proposed ChaBD-BiL model.

The DenseNet-201 construction learns the attributes by utilizing its learnable weights. It is parametrically effectual because of likelihood of feature reuse using diverse layers. Straight links are obtainable from all preceding layers through following layers to indorse connectivity. In order to effectively optimise features, CBAM(Convolutional Based Attention Module), an efficient attention module, infers the attentional map along channel and spatial dimensions. To obtain weighted results, the characteristics are first passed via the channel attention module and then the spatial attention module to obtain the final weighted results. The list of annotations used in Eq. (20) to Eq. (29) are shown below.

- $F$  Feature map
- $C(F)$  Channel attention module
- $S(F)$  Spatial attention module
- $\lambda$  Sigmoid function
- $AP$  Average Pooling function
- $MP$  Maximum Pooling function
- $MLP$  Multi layer perceptron
- $W$  Weight matrix
- $B$  Bias factor

The following is the evaluation formula for the channel and spatial attention modules.

$$C_A(F) = \lambda(MLP(AP(F) + MLP(MP(F)))) \quad (20)$$

$$S_A(F) = \lambda\left(F\left(Concat\left(AP(F), MP(F)\right)\right)\right) \quad (21)$$

From the above expressions,  $F$  indicates the feature map,  $C_A(F)$  denotes the channel attention module and  $S_A(F)$  indicates the spatial attention module,  $\lambda$  represents the sigmoid function,  $AP$  indicates average pooling function,  $MLP$  indicates Multi – Layer Perceptron and  $MP$  represents maximum pooling function. The feature concatenation can be expressed as below.

$$F^I = N_I([F^0, F^1, \dots, F^{I-1}]) \quad (22)$$

From the above expression,  $N_I(\cdot)$  signifies the non-linear transformation that is represented as a composite function including Batch Normalization (BN), ReLU, Convolution and CBAM. The

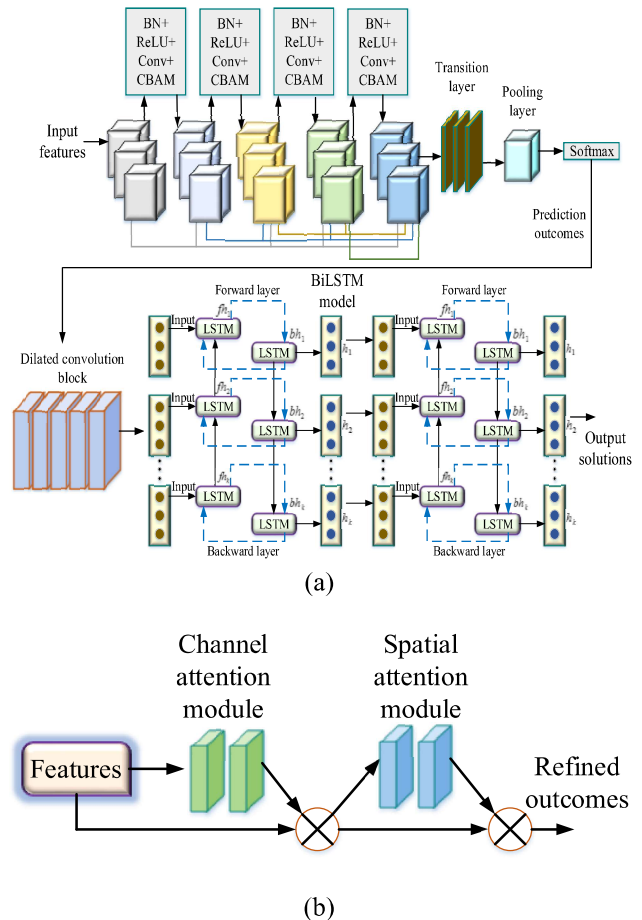


Figure. 4 Model architecture of (a) ChaBD-BiL (b) CBAM module

amalgamation of feature maps equivalent to layer 0 to  $I - 1$  can be designated as  $[F^0, F^1, \dots, F^{I-1}]$ . For the purpose of down sampling, dense blocks comprising of BN, convolutional, ReLU, CBAM and average pooling layers are produced. The pooling layer slowly reduces the size of feature to diminish the parameters and prediction complexity. The grouping of feature maps from dense block is carried out and the dimensions are minimized through the transition layer. Finally, the features can be predicted into good, not bad and poor. From this, the performance outcomes of educational institution can be predicted based on the student review data. Corresponding to the predicted outcomes, the solutions can be recommended using Dilated Convolution BiLSTM to enhance the educational institution performance. The dilated convolution can subjectively increase the receptive field of small convolution kernels to enhance the recommendation accuracy. Without maximizing the number of parameters, the Dilated Convolution BiLSTM can sample the underlying feature maps. The neurons present in the LSTM [30] model comprises of output gate, input gate, forget gate and memory cell. The forget gate is used for data identification from the

previous state  $m_{t-1}$  that is not to be remembered based on current input.

$$d_t = \phi(W_{ud}u_t + W_{vd}m_{t-1} + B_d) \quad (23)$$

From the above expression,  $\phi$  means sigmoid activation function,  $W_{ud}$  indicates the weight matrix between  $u_t$  and  $d_t$ .  $W_{vd}$  signifies the weight matrix between  $m_{t-1}$  and  $d_t$ . The trainer input at time  $t$  is indicated as  $d_t$ , output of previous hidden layer is meant as  $m_{t-1}$  and the bias factor is given as  $B_d$ . Similarly, the input gate can be expressed as follows.

$$e_t = \phi(W_{ue}u_t + W_{ve}m_{t-1} + B_e) \quad (24)$$

The output gate can be mathematically expressed as follows.

$$g_t = \phi(W_{ug}u_t + W_{vg}m_{t-1} + B_g) \quad (25)$$

The final results of LSTM cell are cell output state ( $C_t$ ) and layer output ( $m_t$ ) which can be given as follows.

$$C_t = d_t \otimes C_{t-1} + e_t \otimes \widehat{C}_t \quad (26)$$

$$m_t = g_t \otimes \tanh(C_t) \quad (27)$$

The intermediate cell input state is meant as  $\widehat{C}_t$  and it can be expressed as follows.

$$\widehat{C}_t = \tanh(W_{uc}z_t + W_{vc}m_{t-1} + B_c) \quad (28)$$

As, LSTM cannot use the suitable information, BiLSTM includes both LSTMs that assimilate information from mutual directions. The forward LSTM directs the input from left to right and evaluates the hidden state ( $\vec{m}_t$ ) based on  $z_t$  and  $m_t - 1$ . The backward LSTM directs the input from right to left and examines  $\overleftarrow{m}_t$  hidden state based on  $z_t$  and  $m_t - 1$ . In a BiLSTM network, the forward and backward parameters are unrelated to one another. The final hidden state of BiLSTM model integrating the forward and backward directional vector at time ( $t$ ) can be expressed as follows.

$$m_t = [\vec{m}_t, \overleftarrow{m}_t] \quad (29)$$

Through the proposed ChaBD-BiL model, the solutions like minor improvements are required, no further improvement required and need to improve a

Table 3. Hyper parameter details

Sl. No	Hyper-parameters	Proposed model
1.	Batch size	60
2.	Initial learning rate	0.0001
3.	Learning algorithm	Adam
4.	Maximum epoch size	100
5.	Activation function	ReLU
6.	Maximum iteration	100

lot can be recommended effectively. Based on the recommendation decisions obtained from student review data, the educational institution can promote appropriate actions.

## 5. Results and discussion

The proposed ChaBD-BiL model is explored with varied stages like pre-processing, feature extraction, selection, clustering, and recommendation. The experimental outcomes of the proposed ChaBD-BiL model are signified in this section. The performances of the proposed model are evaluated using PYTHON simulation platform. Various existing approaches are associated with the recommended model to evaluate the performance. The dataset details, description of the performance metrics and its mathematical formulation, performance analysis and analogy are established in the succeeding sections. Combining and changing the different parameters can affect the accuracy of the machine learning algorithms [33]. Table 3 illustrates the hyper parameter setting of suggested model.

### 5.1 Dataset description

Student review dataset is utilized in the proposed model and has been collected from online Kaggle source given as follows <https://www.kaggle.com/datasets/brarajit18/student-feedback-dataset?resource=download>. The dataset is acquired from North India students belongs to a prominent university. The overall institutional report is gathered based on the student feedback data. The dataset includes six categories like course content, teaching, library facilities, examination, lab work and extracurricular activities. In addition to the dataset some attributes like accommodation facilities, hostel food facility, transport facility, cleanliness, canteen, prediction outcomes and recommendation solution are added manually.

Table 4. Performance metrics and its formulation

Performance Metrics	Description	Mathematical formulation
Accuracy	Accuracy can be defined as the total of true positive and false metrics added to the total of true and false metrics.	$A = \frac{W + X}{W + X + Y + Z}$ W -True positive, X -True negative, Y -False positive, Z -False negative.
Kappa	The stability of prediction and employment of probabilistic assessments amongst the predictable scores in case of agreement and disagreement is terms as Kappa Score.	$K = \frac{\lambda_0 - \lambda_f}{1 - \lambda_f}$ $\lambda_0$ - Score agreement between predicted and actual value $\lambda_f$ - Score disagreement between actual and predicted ones.
Specificity	Specificity is the quantity of negative results to the total sample that are actually negative.	$SPE = \frac{X}{X + Y}$ X -True negative, Y -False positive,
F1 score	The combination of precision and recall to a single value is termed an F1 score.	$F1S = 2 * \frac{PPV * TPR}{PPV + TPR}$ PPV - Positive predictive value TPR -True positive rate
Sensitivity	Recommendation outcomes are highly sensitive if the data produces positive cases.	$SEN = \frac{W}{W + Z}$ W -True positive, Z -False negative.
MAE	The prediction error between predicted and actual outcomes is called MAE.	$MAE = \frac{\sum_{p=1}^M  x_p - y_p }{M}$ x -Predicted value, y -Actual value M-Total samples
RMSE	RMSE designates the standard deviation of recommendation errors.	$RMSE = \sqrt{\frac{\sum_{p=1}^M (x_p - y_p)^2}{M}}$ x -Predicted value, y -Actual value M-Total samples

### 5.2 Performance metrics

The proposed recommendation model can be evaluated on considering diverse metrics like Accuracy, Sensitivity, Specificity, F1 score, Kappa, mean absolute error (MAE) and Root mean square error (RMSE). The performance metrics are described with its mathematical formulations for examining the proposed performance in Table 4.

### 5.3 Baseline model comparison analysis

The Proposed model is associated with several existing approaches to prove the superiority of the proposed approach. The existing methodologies like auto encoder (AE), deep convolutional neural network (DCNN), bidirectional gated recurrent unit (BiGRU) and BiLSTM are considered for comparison. The performance outcomes in terms of diverse evaluated metrics like Accuracy, Sensitivity,

Table 5. Performance comparison analysis

Performance outcomes (%)	Techniques				
	AE	DCNN	BiGRU	BiLSTM	Proposed
Accuracy	88.10	89.90	91.71	93.15	98.19
Sensitivity	78.75	87.14	89.00	90.63	97.41
Specificity	79.97	89.58	90.45	92.66	98.25
F1 score	68.14	72.48	74.49	76.52	91.28
Kappa	44.59	71.17	71.80	75.33	91.28
MAE	0.18	0.16	0.13	0.11	0.03
RMSE	0.42	0.43	0.37	0.37	0.21

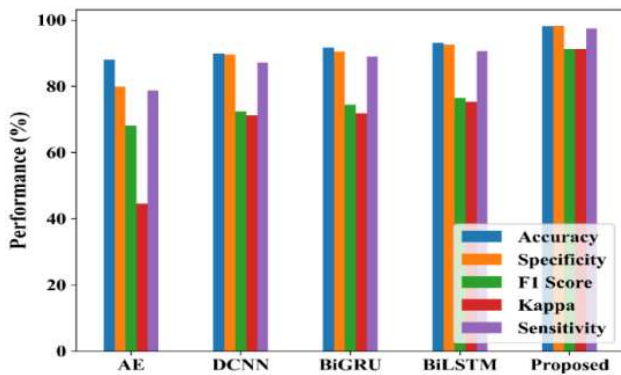


Figure. 5 Performance Comparison of recommendation models

Specificity, F1 score, kappa, MAE and RMSE are showed in Table 5.

From the below table, it can be obviously analysed that the proposed model obtained better performance outcomes compared to the existing approaches. The performance of the proposed model is analyzed by comparing to the existing models like AE, DCNN, BiGRU and BiLSTM.

The above graphical represents clearly that the accuracy of proposed ChaBD-BiL model in solution recommendation based on the student review data is 98.19%, sensitivity as 97.41%, specificity as 98.25% and F1 score as 91.28%. Better accuracy rate can be obtained by focussing over the most appropriate features through effectual procedures for processing text data. Improved training ability and only slight errors are perceived by handling optimal features. The existing models like AE, DCNN, BiGRU and BiLSTM has accomplished less performance than the proposed model because of certain drawbacks like huge accumulation of features, high testing time, less convergence and less feature learning capability. Figure 6 (a)-(b) indicates the performance attained by the proposed and existing techniques in terms of MAE and RMSE.

For an enhanced recommendation model, the MAE and RMSE value must be less. The MAE and

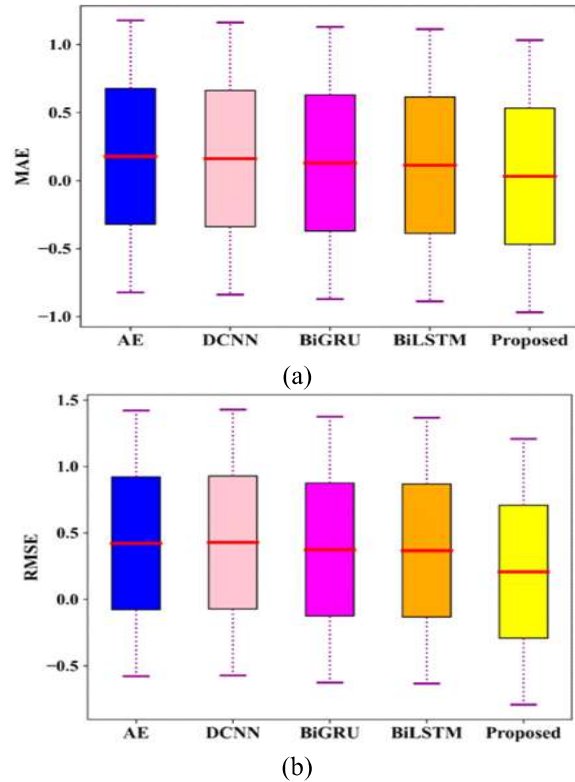


Figure. 6 Error performance analysis: (a) MAE and (b) RMSE

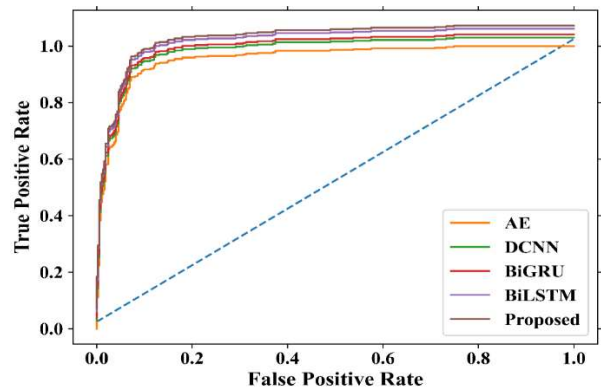


Figure. 7 ROC analysis of Existing & Proposed models

RMSE performance of proposed ChaBD-BiL model and existing models are analysed.

From the above graphical representation, the proposed RMSE is attained to be 0.21 and MAE as 0.03 respectively. When compared to the MAE and RMSE value of proposed model, existing models attained increased error rates. Because of the use of incapable features, the existing models are highly prone to increased error rate. Hence, it can be justified that the proposed model has obtained lesser rates of MAE and RMSE. Fig. 7 illustrates the ROC curve analysis of proposed and existing models.

The ROC curve is examined in terms of false positive rate (FPR) and True positive rate (TPR). The optimum cut-off depicts the supreme TPR or

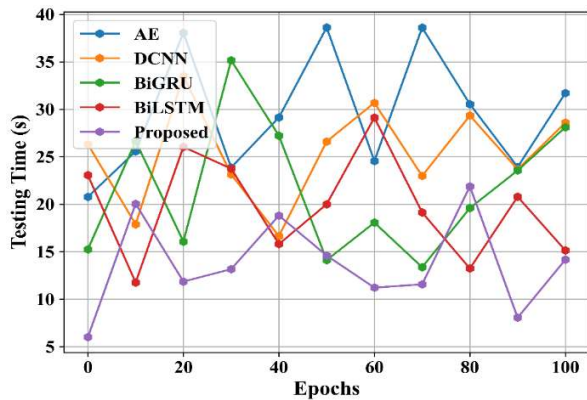


Figure. 8 Testing time analysis

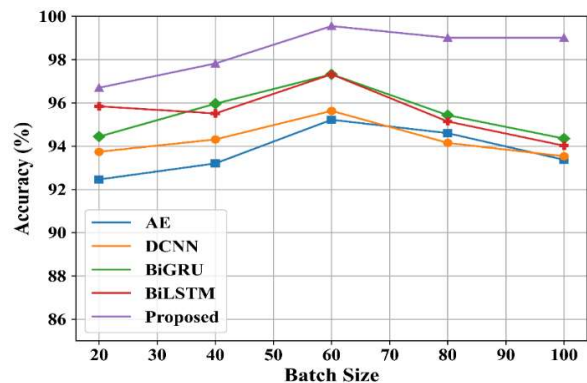


Figure. 9 Accuracy Vs Batch size performance analysis

sensitivity with least FPR or specificity. The ROC curves are often investigated to expose the trade-off between TPR and FPR for every probability. It designates the efficacy of recommendation model and it denotes the degree of ability in predicting the feedback performance. Higher the rate of ROC indicates better the performance of recommendation model. Fig. 8 depicts the testing time analysis of Proposed and Existing methods in terms of student feedback dataset.

When comparing the testing time of proposed ChaBD-BiL model with existing approaches, the proposed testing time is highly lesser than the existing methods like AE, DCNN, BiGRU and BiLSTM. The testing time performance is analysed by varying the epoch size from 0 to 100. The proposed recommendation model has attained 14.17 seconds for testing whereas the existing AE obtained 31.71 seconds,

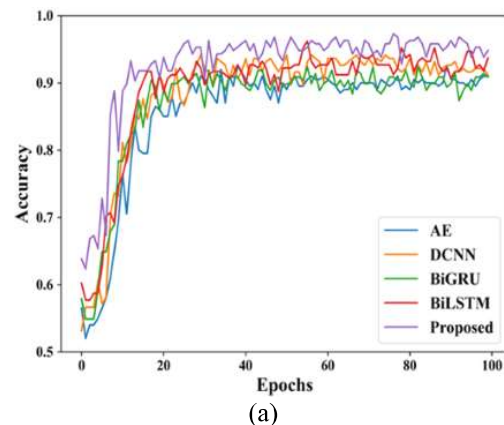
DCNN as 28.59 seconds, BiGRU as 28.11 seconds and BiLSTM as 15.14 seconds respectively. Because of huge accumulation of features, degraded learning ability and less convergence, existing approaches obtained enhanced testing time. Through this analysis, *It is clearly evident that the proposed algorithm offers a better performance.* Fig.9 indicates the accuracy performance by varying the training batch sizes.

The batch size is denoted as the amount of training data required for single iteration. The suggested model analyses the performance of accuracy under varying batch sizes and, the obtained outcome is compared with different existing techniques. The performance of proposed model is analysed by varying the batch size from 20 to 100 whereas higher accuracy is attained when the training batch size is set to be 60.

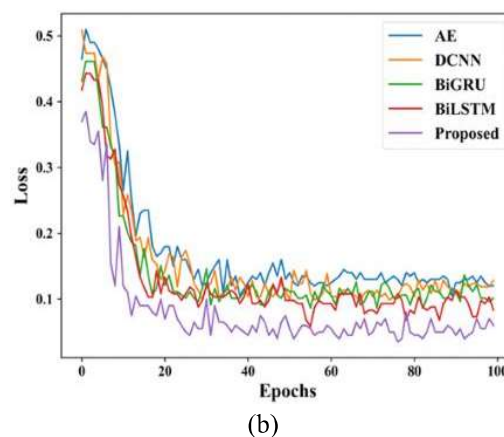
#### 5.4 Accuracy and loss evaluation measures

The accuracy and loss of the proposed ChaBD-BiL model for solution recommendation are analysed with testing data. In the proposed research work, 80% of data is used for model training and 20% is utilized for model testing. The accuracy and loss performance obtained during testing stages in processing student review data are provided as follows. Figure 10 (a)-(b) indicates the testing performance analysis in terms of accuracy and loss.

To evaluate the learning performance of the commended ChaBD-BiL model, the Accuracy and loss curves are analysed. The accuracy and loss under testing phases are assessed by varying the epoch size from 1 to 100 consecutively.



(a)



(b)

Figure. 10 Testing curve analysis: (a) Accuracy and (b) Loss

Due to increased epoch size, increase in accuracy and decrease in loss may happen. The existing models like AE, DCNN, BiGRU and BiLSTM are analysed with respect to testing data. The testing phase of these existing methods were found to be slower than the proposed model and so reaching of greater accuracy is complicated. The existing models consumes more time for testing and so the computational time tends to be high. In the loss curve, the proposed testing loss gets diminished in case of increased epoch size. When compared with the existing architectures, the proposed recommendation model obtains higher accuracy with condensed loss. High losses are attained in the existing approaches because of increased time complexity, degraded training ability and convergence issues.

## 6. Conclusion

In the proposed research work, precise recommendation of solutions can be obtained based on the student feedback data to enhance the educational institution performance using novel approaches. Here, student feedback data was collected from Kaggle source and some of the attributes were added manually. The text data was pre-processed using Stemming, Tokenization, case - folding and stop - word - removal procedures. Effective features were extracted using ELexBert model and the most significant features were selected using ACoaT algorithm. The selected features were clustered using Uden\_KMC model based on feature similarity. A novel hybrid DL based ChaBD-BiL model was employed for recommending better decisions. The drawbacks like degraded training ability, overfitting, time consumption and convergence issues were overcome through efficient learning of input data. The recommendation performances are analysed using PYTHON simulation platform whereas the overall accuracy of 98.19%, specificity of 98.25%, F1 score of 91.28%, sensitivity of 97.41% and Kappa score of 91.28% are obtained. Lesser rates of MAE as 0.03 and RMSE as 0.21 were obtained due to effective utilization of optimal features. Also, the testing time was less in recommending an appropriate solution for the input data. In future, the proposed work can be extended further with the utilization of larger datasets. Also, the consideration of features will be more optimal with the adoption of enhanced hybrid optimization strategies to enhance the recommendation accuracy.

## Conflicts of Interest

The authors have no conflicts of interest to declare. All co-authors have seen and agreed with the contents of the manuscript and there is no financial interest to report. We certify that the submission is original work and is not under review at any other publication.

## Author Contributions

Conceptualization, methodology, software, validation, formal analysis, writing-original paper draft, Naga Jyothi; writing-review and editing, Dr. Uma Dulhare; visualization, Naga Jyothi; supervision, Dr. Uma Dulhare.

## References

- [1] S. Pratsri, P. Nilsook, and P. Wannapiroon, "Synthesis of data science competency for higher education students", *International Journal of Education and Information Technologies*, Vol. 16, pp.101-109, 2022.
- [2] D. Carless, and N. Winstone, "Teacher feedback literacy and its interplay with student feedback literacy", *Teaching in Higher Education*, Vol. 28, No.1, pp.150-163, 2023.
- [3] D. Carless, "From teacher transmission of information to student feedback literacy: Activating the learner role in feedback processes", *Active Learning in Higher Education*, Vol. 23, No. 2, pp.143-153, 2022.
- [4] R.A. de Kleijn, "Supporting student and teacher feedback literacy: an instructional model for student feedback processes", *Assessment & Evaluation in Higher Education*, Vol. 48, No. 2, pp.186-200, 2023.
- [5] K. Sangeetha, and D. Prabha, "RETRACTED ARTICLE: Sentiment analysis of student feedback using multi-head attention fusion model of word and context embedding for LSTM", *Journal of Ambient Intelligence and Humanized Computing*, Vol.12, No. 3, pp.4117-4126, 2021.
- [6] D. Baneres, M.E. Rodríguez-Gonzalez, and M. Serra, "An early feedback prediction system for learners at-risk within a first-year higher education course", *IEEE Transactions on Learning Technologies*, Vol. 12, No. 2, pp.249-263, 2019.
- [7] K.D. Vattøy, and K. Smith, "Students' perceptions of teachers' feedback practice in teaching English as a foreign language", *Teaching and Teacher Education*, Vol. 85, pp.260-268, 2019.

- [8] I.A. Chounta, K. Uiboleht, K. Roosimäe, M. Pedaste, and A. Valk, "From data to intervention: predicting students at-risk in a higher education institution", In: *Companion proc. 10th international conference on learning analytics & knowledge (LAK20)*, 2020.
- [9] A.A. Mubarak, H. Cao, and W. Zhang, "Prediction of students' early dropout based on their interaction logs in online learning environment", *Interactive Learning Environments*, Vol. 30, No. 8, pp.1414-1433, 2022.
- [10] S. Sukmawati, S. Sujarwo, D. N. Soepriadi, and N. Amaliah, "Online English Language Teaching in the Midst of Covid-19 Pandemic: Non EFL Students' Feedback and Response", *Al-Ta lim Journal*, Vol. 29, No. 1, pp.62-69, 2022.
- [11] E. Ossiannilsson, K. Williams, A. F. Camilleri, and M. Brown, "Quality models in online and open education around the globe. State of the art and recommendations", *Oslo: International Council for Open and Distance Education*, 2015.
- [12] R. Ammigan, "Institutional satisfaction and recommendation: What really matters to international students", *Journal of International Students*, Vol. 9, No.1, pp.262-281, 2019.
- [13] D. Nicol, D, "The power of internal feedback: Exploiting natural comparison processes", *Assessment & Evaluation in higher education*, Vol. 46, No. 5, pp.756-778, 2021.
- [14] Y. Zhu, H. Lu, P. Qiu, K. Shi, J. Chambua, and Z. Niu, "Heterogeneous teaching evaluation network based offline course recommendation with graph learning and tensor factorization", *Neurocomputing*, Vol. 415, pp.84-95, 2020.
- [15] T.R. Guskey, "Grades versus comments: Research on student feedback", *Phi Delta Kappan*, Vol. 101, No.3, pp.42-47, 2019.
- [16] H. De Wit, and P. G. Altbach, "Internationalization in higher education: global trends and recommendations for its future", *Higher education in the next decade*, pp. 303-325, 2021.
- [17] F.J. García-Peñalvo, A. Corell, V. Abella-García, and M. Grande-de-Prado, "Recommendations for mandatory online assessment in higher education during the COVID-19 pandemic", In: *Radical solutions for education in a crisis context: COVID-19 as an opportunity for global learning*, pp. 85-98, 2020.
- [18] K. Okoye, A. Arrona-Palacios, C. Camacho-Zuñiga, J.A.G. Achem, J. Escamilla, and S. Hosseini, "Towards teaching analytics: a contextual model for analysis of students' evaluation of teaching through text mining and machine learning classification", *Education and Information Technologies*, pp.1-43, 2022.
- [19] M.A. Haque, D. Sonal, S. Haque, M. Rahman, and K. Kumar, "Learning management system empowered by machine learning", In: *AIP Conference Proceedings*, Vol. 2393, No.1, AIP Publishing, 2022.
- [20] F.G. Karaoglan Yilmaz, and R. Yilmaz, "Student opinions about personalized recommendation and feedback based on learning analytics", *Technology, knowledge and learning*, Vol. 25, pp.753-768, 2020.
- [21] S. Sood and M. Saini, "Hybridization of cluster-based LDA and ANN for student performance prediction and comments evaluation", *Education and Information Technologies*, Vol. 26, No.3, pp.2863-2878, 2021.
- [22] Z. Yangsheng, "An AI based design of student performance prediction and evaluation system in college physical education", *Journal of Intelligent & Fuzzy Systems*, Vol.40, No.2, pp.3271-3279, 2021.
- [23] Z. Kanetaki, C. Stergiou, G. Bekas, C. Troussas and C. Sgouropoulou, "A hybrid machine learning model for grade prediction in online engineering education", *Int. J. Eng. Pedagog*, Vol. 12, No.3, pp.4-23, 2022.
- [24] F. Ouyang, M. Wu, L. Zheng, L. Zhang and P. Jiao, "Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course", *International Journal of Educational Technology in Higher Education*, Vol. 20, No. 1, pp.4, 2023.
- [25] M. Panda, "Developing an efficient text pre-processing method with sparse generative Naive Bayes for text mining", *International Journal of Modern Education and Computer Science*, Vol. 10, No.9, pp.11, 2018.
- [26] A. Subakti, H. Murfi, and N. Hariadi, "The performance of BERT as data representation of text clustering", *Journal of big Data*, Vol. 9, No.1, pp.15, 2022.
- [27] M. Dehghani, Z. Montazeri, E. Trojovská, and P. Trojovský, "Coati Optimization Algorithm: A new bio-inspired metaheuristic algorithm for solving optimization problems", *Knowledge-Based Systems*, Vol. 259, pp.110011, 2023.
- [28] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants



analysis, and advances in the era of big data”, *Information Sciences*, Vol. 622, pp.178-210, 2023.

- [29] S. Santhanam, “Context based text-generation using lstm networks”, *arXiv preprint arXiv:2005.00048*, 2020.
- [30] U.N. Dulhare, D.N. Jyothi, B. Balimidi, and R. R. Kesaraju, “Classification Models in Education Domain Using PSO, ABC, and A2BC Metaheuristic Algorithm-Based Feature Selection and Optimization”, *Machine Learning and Metaheuristics: Methods and Analysis*, pp. 255-270, 2023.
- [31] U.N. Dulhare, and S. Gouse, “Hands on MAHOUT—machine learning tool”, *Machine Learning and Big Data: Concepts, Algorithms, Tools and Applications*, pp.361-421, 2020.
- [32] F. Arif, F. and U.N. Dulhare, “A machine learning based approach for opinion mining on social network data”, In: *Computer Communication, Networking and Internet Security: Proceedings of IC3T 2016*, pp. 135-147, 2017.
- [33] U.N. Dulhare, and E. H. Houssein, editors, “Machine Learning and Metaheuristics: Methods and Analysis”, *Springer Nature*, 2023.
- [34] U. N. Dulhare, “Prediction system for heart disease using Naive Bayes and particle swarm optimization”, *Biomedical Research*, Vol. 29, No.12, pp.2646-2649, 2018.
- [35] A. Mubeen, and U. N. Dulhare, "Metaheuristic Algorithms for the Classification and Prediction of Skin Lesions: A Comprehensive Review”, *Machine Learning and Metaheuristics: Methods and Analysis*, pp.107-137, 2023.
- [36] U. N. Dulhare, and S.T.A. Taj, “Water quality risk analysis for sustainable smart water supply using adaptive frequency and BiLSTM”, In: *Proc. of International virtual conference on industry*, pp. 67-82, 2021.
- [37] B. Arathi, and U.N. Dulhare, ” Classification of cotton leaf diseases using transfer learning-DenseNet-121”, In: *Proc. of third international conference on advances in computer engineering and communication systems: ICACECS 2022*, pp. 393-405, 2023.
- [38] P.D. Kusuma and A. Dinimaharawati, “Extended stochastic coati optimizer”, *International Journal of Intelligent Engineering and Systems*, Vol. 16, No. 3, pp.482-494, 2023, doi: 10.22266/ijies2023.0630.38.
- [39] P.D. Kusuma and A. Novianty, “Total Interaction Algorithm: A Metaheuristic in which Each Agent Interacts with All Other

Agents”, *International Journal of Intelligent Engineering & Systems*, Vol. 16, No. 1, 2023, doi: 10.22266/ijies2023.0228.20.